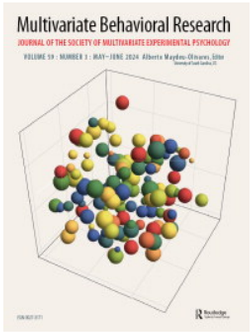

Multivariate Behavioral Research

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hmbr20

Using Monte Carlo Simulation to Forecast the Scientific Utility of Psychological App Studies: A Tutorial

Sebastian Kueppers, Richard Rau & Florian Scharf

To cite this article: Sebastian Kueppers, Richard Rau & Florian Scharf (11 Jul 2024): Using Monte Carlo Simulation to Forecast the Scientific Utility of Psychological App Studies: A Tutorial, Multivariate Behavioral Research, DOI: 10.1080/00273171.2024.2335411

To link to this article: https://doi.org/10.1080/00273171.2024.2335411

Published online: 11 Jul 2024.

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2024.2335411

Routledge
Taylor & Francis Group

Using Monte Carlo Simulation to Forecast the Scientific Utility of Psychological App Studies: A Tutorial

Sebastian Kueppers[a,b,c], Richard Rau[a,b], and Florian Scharf[d]

[a]University of Münster; [b]Institute for Mind, Brain and Behavior, HMU Health and Medical University Potsdam, Germany; [c]University of Hamburg; [d]University of Kassel

ABSTRACT

Mobile applications offer a wide range of opportunities for psychological data collection, such as increased ecological validity and greater acceptance by participants compared to traditional laboratory studies. However, app-based psychological data also pose data-analytic challenges because of the complexities introduced by missingness and interdependence of observations. Consequently, researchers must weigh the advantages and disadvantages of app-based data collection to decide on the scientific utility of their proposed app study. For instance, some studies might only be worthwhile if they provide adequate statistical power. However, the complexity of app data forestalls the use of simple analytic formulas to estimate properties such as power. In this paper, we demonstrate how Monte Carlo simulations can be used to investigate the impact of app usage behavior on the utility of app-based psychological data. We introduce a set of questions to guide simulation implementation and showcase how we answered them for the simulation in the context of the guessing game app *Who Knows* (Rau et al., 2023). Finally, we give a brief overview of the simulation results and the conclusions we have drawn from them for real-world data generation. Our results can serve as an example of how to use a simulation approach for planning real-world app-based data collection.

KEYWORDS

App-based data collection; Monte Carlo simulations; dropout; longitudinal data; planned missingness

Introduction

Mobile applications are frequently used for psychological data assessment in various fields, such as treatment of depression symptoms (Roepke et al., 2015), cognitive screening for dementia of the elderly (Brouillette et al., 2013; Zorluoglu et al., 2015), or the relationship between alcohol consumption and risk-taking tendency (Smith et al., 2017). They are advantageous with respect to external or ecological validity (Harari et al., 2016) and evoke high acceptance among subjects (Ben-Zeev et al., 2014; Miner et al., 2016). Moreover, they offer the possibility to provide the participants with personalized feedback (Wenz et al., 2022), to notify them whenever they need to engage in data collection again (Alkhaldi et al., 2016; Pavliscsak et al., 2016), and to transfer data in real time (Fischer & Kleen, 2021). They also allow for an easy implementation of planned missingness designs, for example by presenting only a random subset of stimuli to each subject. However, aside from being time-consuming and costly to develop, mobile applications also offer less control over the participation process compared to traditional on-site, offline studies which threatens data quality. For example, Torous et al. (2020) estimated in their meta-analysis that the dropout rate in studies that used apps to collect depressive symptoms was nearly 50%.

Consequently, researchers are confronted with the challenge to weigh the benefits offered by app-based data assessment against the disadvantages in terms of data quality. To accomplish this, they must know the type and amount of data their app would need to produce in order to be useful for answering their substantive research question.

Other than for most common study designs such as randomized experiments or cross-sectional surveys, however, there is no straightforward way to calculate criteria for scientific utility, including the power or required sample size of an app study. The reason for this is that psychological data assessed with an app is

CONTACT Florian Scharf florian.scharf@uni-kassel.de University of Kassel, Kassel, Germany.

characterized by possibly complex data structures with high degrees of data missingness due to dropouts and complex missingness patterns. Explicit formulas for analyses such as power calculations are not available for these data structures. Therefore, it is challenging to plan the required sample in advance in the context of app-based studies. Instead of disregarding these challenges or avoiding the endeavor altogether, we assert that a prudent approach would involve stepping back and systematically examining the conditions under which the proposed app study can yield actual scientific insights.

In the present article, we introduce Monte Carlo simulation studies as solutions to these challenges. Monte Carlo simulation is a statistical technique where sample data are generated repeatedly from a known population model and are then analyzed in each sample. Properties of these analyses such as estimation bias or statistical power are then examined by averaging the results across all samples (Morris et al., 2019). This technique allows researchers to manipulate all data-generating mechanisms and thereby explore how different scenarios (e.g., different sample sizes, dropout rates, confounding influences, and so forth) would impact their outcome metric of interest. In many research settings, this outcome metric is the statistical power when testing a particular hypothesis. Other outcome metrics may include the relative bias in the estimation of a particular fixed effect or, as showcased in this paper, the reliability of random effects estimated from the data. Throughout this article, we will use the term "scientific utility" to encapsulate all criteria researchers might find useful to evaluate whether their research question can be effectively addressed using app data. We aim to provide guidelines for Monte Carlo simulations of app-based psychological data. For this matter, we will outline considerations researchers need to make when planning such a simulation study and demonstrate them using the simulation we conducted in the context of the guessing game app *Who Knows* (Rau et al., 2023).

Background

One major challenge for app-based psychological data assessment is that the longitudinal nature of the underlying research design may complicate sample size considerations. For example, Roepke et al. (2015) evaluated the effects of an app-based clinical treatment over time. They tracked the progress of patients who used their depression and anxiety treatment application *SuperBetter*. Patients were to use the app over a period of time and the potential reduction in symptoms was investigated to evaluate the app's efficacy. A problem in such setups is that all variables of theoretical interest might change over time, but collecting many time-variant covariates can be burdensome—potentially reducing the participants' compliance. Furthermore, since data for a single subject is collected at multiple times, longitudinal data are statistically dependent—but this dependency is also a function of the interval between measurements. And this may not be the only data dependency to account for. For instance, if Roepke et al. (2015) would have provided therapeutic assistance to their participants, data might systematically vary across therapists, too, representing another source of interdependence. A similar logic can be applied when random samples of stimuli or items are presented to participants making them sources of variance that need to be accounted for (Judd et al., 2012; Yarkoni, 2022).

In addition, longitudinal data practically always display missingness to some degree. This might be because subjects drop out of the study and thus do not provide any data beyond a certain measurement time or only respond to a subset of study variables or at a subset of the measurement occasions. Consequently, app-based psychological data assessment is generally characterized by a high degree of dropouts (see, e.g., Meyerowitz-Katz et al., 2020; Torous et al., 2020) and missing data, too. As a result of a highly unbalanced number of observations across participants, estimates of state variables (e.g., average mood levels of participants) may differ in their reliability. This presents another challenge for data analysis, where results are biased unless missingness is handled appropriately.

A related challenge for researchers planning an app-based study is that the dropout rate in an app study is difficult to predict. In Meyerowitz-Katz et al. (2020) meta-analysis of app studies on mental health, the dropout rates of observational studies varied between 91% in a 1-year lasting RCT and 18% in an observational trial that lasted 6 weeks. They point out, however, that their findings are "limited by high heterogeneity and the lack of reporting in many trials on attrition rates" (Meyerowitz-Katz et al., 2020, p. 8). According to Pfammatter et al. (2017), there are some aspects that can explain high dropout rates, such as misunderstanding of the trial's components, low perceived usefulness of the app, or little recall of the incentive structure. In summary, app-based psychological data always displays missingness, although the exact extent and pattern is difficult to predict.

The Who Knows application

Who Knows (Rau et al., 2023) is a mobile application that aims to gather psychological data in an affordable and joyful way that minimizes participant burden, facilitates participation, and yields high ecological validity. Its data can address a myriad of questions within person perception research and beyond, including the psychological characteristics of the "good judge" (a person who judges others more accurately than most other judges), the "good target" (a person who is judged by others more accurately than most other targets), and the "good trait" (a characteristic that is judged more accurately than most other characteristics; Funder, 2012).

Since participation is not externally incentivized, any user who downloads the app—who we call *perceiver* from now on– can partake in as many game rounds as they like. When starting a game in the application, the perceiver is shown a short introductory video of a target person. In this video, the target provides a brief self-introduction of themselves including their profession, hobbies, relationship status, and three adjectives which describe them. After watching the video, the perceiver is asked to judge the target in terms of five everyday characteristics. Example items are "Has [the target] ever experienced heartbreak?" or "Has [the target] ever been involved in politics?". For each accurate judgment, i.e., for each response that matches the target's self-description on an item, the perceiver acquires points. After each game, the perceiver receives feedback about their performance. In the app's feedback area, they can also get an overview of their overall performance compared to other users.

The target who is presented to the perceiver in a game is selected randomly from a pool of 50 available targets. However, as a constraint by the app algorithm, each target is to be rated once (twice, thrice, ...) before they can be rated a second (third, fourth, ...) time by a given perceiver. For each target, there exists a pool of roughly 60 items on which they have provided self-reports as accuracy criteria. These items are a random subset out of a pool of 820 items in total. The five items presented in a specific game are sampled from all items which the perceiver has never answered for the presented target before. That is, every combination of a perceiver, a target, and a specific item can occur at most once in the dataset.

Like the data from other psychological apps, *Who Knows* data have a hierarchical structure, as there are multiple observations for each perceiver, target, and item. This implies a cross-classified multilevel structure with observations on a lower level and perceivers, targets, and items as higher-level units. This form of interdependency among different datapoints needs to be accounted for in the analysis of *Who Knows* data.

Furthermore, the *Who Knows* application can be expected to yield considerable dropout rates (i.e., participants who stop playing after some games) and a substantial amount of missing data. In fact, we expected exceptionally high levels of missingness because our approach to rely exclusively on intrinsic motivation implies that perceivers need not judge every target on every available item to contribute to the data collection. Outcomes for other combinations of targets and items for these perceivers would be missing. In addition, a high degree of data combinations is expected to be missing *a priori* because the present targets gave self-report only regarding a random subset of items, resulting in a source of *planned missingness* in the data (Graham et al., 2006).

As a result, it was questionable for us whether *Who Knows* could produce enough data to make up for the deficits in data quality that result from the voluntary participation approach. Traditional power analyses fell short due to the incompleteness and complexity of the data. Also, given that the collection of person perception data through a non-incentivized guessing game is unprecedented in the literature, we were essentially agnostic about the amount of missing data that was to be expected. Consequently, we decided to simulate *Who Knows* data under varying degrees of missingness to evaluate the app's scientific utility, as a function of missingness. Our simulation script can be retrieved from tinyurl.com/appstudysimulation. (Rau et al., 2023)

Tutorial

To prepare a Monte Carlo simulation, one must first answer several questions regarding the underlying statistical model, study design, anticipated missingness patterns, and data analysis. In the following, we will outline these necessary considerations. We will describe which questions must be asked regarding each aspect, outline their general relevance and implications for the simulation and, finally, illustrate which decisions were made for the simulation of the *Who Knows* data.

We conducted our simulation using the statistics software *R*, version 4.2.1 (R Core Team, 2022). Many *R* packages are available that allow for a user-friendly implementation of simulation studies by providing population parameters for different population models. Examples are *simsem* (Pornprasertmanit et al., 2021) for structural equation models, *SIMR* (Green & MacLeod, 2016) for (G)LMMs, or *powRICLPM* (Mulder, 2022) for data simulations with random intercept cross-logged panel modeling. Other statistical software packages such as Mplus (Muthén & Muthén, 1998-2017) also provide simulation functionalities for such purposes.

Data generation process

As a first step, one needs to decide on a generation process behind the data. That is, the properties of the outcome and predictor variables as well as their mathematical relationship need to be specified. To obtain realistic app data, we suggest simulating observations at a level that allows for the highest flexibility in data generation with respect to the population model and the introduction of missingness to the data. Hence, typically data for each observation (e.g., measurement occasion) should be simulated. That is, values for all relevant dependent (outcome) and independent (predictor) variables need to be simulated so that missingness patterns can be flexibly introduced to the data by deletion from a hypothetical complete data set.

What is the outcome variable?

The characteristics of the outcome variable are important for the choice of the statistical model, especially its level of measurement and its distribution. It is also important to think about whether the outcome variable is measured only once or repeatedly. Another aspect to consider is whether the outcome is a manifest or latent variable, and the outcome's unit. Table 1 provides an overview of how typical outcome variables in psychological research can be characterized in terms of the above criteria.

The outcome variable in *Who Knows* is a naturally dichotomous and manifest single-item measure depicting either a perceiver's correct (i.e., equal to 1) or incorrect (i.e., equal to 0) response to a single target on a single item. (Note that there is also a small portion of items which use a continuous response format in the actual app but our simulation focused on the dichotomous case for the sake of simplicity).

Table 1. Examples of outcome variables.

Variable	Level of measurement	Distribution	Latent vs. manifest	Unit
Reaction time	Ratio	e.g., Ex-gaussian	Manifest	(Milli-)seconds
Extraversion	Interval	Gaussian	Latent	Likert scale
Amount of waste of a household	Ratio	Gaussian	Manifest	Kilograms/pounds
Smartphone Ownership	Dichotomous	Bernoulli (Binary)	Manifest	Binary (Yes/No)
Number of Purchases Online	Count	Poisson	Manifest	Count (e.g., per month)

What is the population model?

Apart from these basic characteristics of the outcome variable, it is important for the definition of the population model to identify potential predictor variables and their relationship to the outcome. All predictor variables and factors should be characterized as realistically as possible as well. This may include the independent variables related to the study design that are systematically manipulated as part of the study, such as the study condition or the distance between measurement occasions, on the one hand. On the other hand, there might be covariates or demographic features of the participants that relate to the outcome, like age or income. When assessing longitudinal data, in addition to such time-invariant predictors, there might be time-variant predictor variables. An example of such would be a symptom catalog that is collected in parallel with the actual outcome. As for the outcome variable, all these variables should be described regarding their level of measurement, distribution, and presence of measurement error. For the sake of simplicity, we focus on manifest (vs. latent) variable models during the tutorial part. These models are simpler, because their analysis necessitates less theoretical assumptions and less sophisticated statistical techniques. This also fits with Who Knows as an example case, where data analysis exclusively pertains to manifest variables.

In addition, the specific mathematical relationship between the outcome and the predictor variables must be considered. When choosing a statistical framework to simulate data, several things need to be considered. Repeated measurements, which is common in app-based studies, result in a nested data structure in which responses from the same participant are more similar than responses from different participants (e.g., Hedeker & Gibbons, 2006). In these contexts, it is important that the population model is capable of producing data that mimic these interdependencies in a realistic fashion. To account for this, for instance, linear growth models, generalized linear mixed models (GLMMs), or cross-lagged panel models could be used as population models. In order to evaluate estimation stability under varying usage scenarios, the population model should match the analysis model.

In the following, we will focus on the use of GLMMs as a framework for data synthesis and analysis. GLMMs allow for flexible simulation and analysis of data with specific dependencies or with missingness, as well as for a distinction between fixed and random effects. In addition, GLMMs allow for modeling various types of outcome measures (e.g., continuous, dichotomous or count variables) as well as non-linear relationships between outcome and predictors. Therefore, we consider GLMMs a suitable framework for the simulation of longitudinal data as they are often collected with apps. Another advantage of GLMMs for data simulation is the possibility to accommodate for a range of data with varying degrees of structural complexity such as models with higher-level predictors, models with more than two levels, and models with more complex structures such as crossed random effects. For instance, it has been argued that participants and items should always be considered as (crossed) random effects whenever participants are presented with a set of multiple stimuli (Judd et al., 2012; Yarkoni, 2022). We briefly illustrate the specification of an LMM with random effects across participants using the study of Roepke et al. (2015).

Roepke et al. (2015) were interested in digitally monitoring the efficacy of their depression and anxiety treatment application *SuperBetter*.

In their study, they distinguished between two different versions of the app—which we label *SB1* and *SB2*– and a patient control group, the waiting list, that did not use the app.

In one part of their study, they modeled the level-1 within-person change in the total depression score over time T and predicted the level-1 slope with the dichotomous higher-level study group variables G_{SB1} and G_{SB2}. Because Y_{ij}, the single ith response of patient j, is a continuous variable, the data could be analyzed with the following linear mixed model (LMM):

$$Y_{ij} = \underbrace{c_{00}}_{\text{Fixed intercept}} + \underbrace{u_{0j}}_{\text{Random intercept across participants}} + \left(\underbrace{c_{10}}_{\text{Fixed slope}} + \underbrace{c_{11,SB1}}_{\text{Random slope for study group 1}} \times \underbrace{G_{SB1}}_{\text{Dichotomous variable for study group 1}} + \underbrace{c_{11,SB2}}_{\text{Random slope for study group 2}} \times \underbrace{G_{SB2}}_{\text{Dichotomous variable for study group 2}} + \underbrace{u_{0j}}_{\text{Random slope across participants}} \right) \times \underbrace{T_i}_{\text{Measuring point } i} + \underbrace{\varepsilon_{ij}}_{\text{Level 1 residuum}} \quad (1)$$

The effects of interest in the original study were the differences between the two versions of *SuperBetter* and the control group. In the above model, this would correspond to the parameters of the cross-level interactions $c_{11,SB1}$ and $c_{11,SB2}$. To further illustrate the flexibility of GLMMs, we expand the above example and assume that participants were assigned to different therapists who supported them in using the app. One therapist k supports various participants which results in therapists resembling a higher-order random effect in which participants and observations are nested. Adding therapists as random effect to Equation (1) results in:

$$Y_{ijk} = \underbrace{c_{000}}_{\text{Fixed intercept}} + \underbrace{u_{0j0}}_{\text{Random intercept across participants}} + \underbrace{\boldsymbol{u_{00k}}}_{\text{Random intercept across therapists}} + \left(\underbrace{c_{100}}_{\text{Fixed slope}} + \underbrace{c_{110,SB1}}_{\text{Random slope for study group 1}} \times \underbrace{G_{SB1}}_{\text{Dichotomous variable for study group 1}} + \underbrace{c_{110,SB2}}_{\text{Random slope for study group 2}} \times \underbrace{G_{SB2}}_{\text{Dichotomous variable for study group 2}} + \underbrace{u_{1j0}}_{\text{Random slope across participants}} + \underbrace{\boldsymbol{u_{10k}}}_{\text{Random slope across therapists}} \right) \times \underbrace{T_i}_{\text{Measuring point } i} + \underbrace{\varepsilon_{ijk}}_{\text{Level 1 residuum}} \quad (2)$$

The random therapist intercept u_{00k} depicts the effect therapist k has on depression symptoms at the beginning of the study for $T_i = 0$. The random therapist slope effect u_{10k} resembles the effect therapist k has on the development of depression symptoms over time.

We believe that investigation of psychological symptoms over time, like Roepke et al. (2015) did, is a typical use case of an app in psychological research. The resulting longitudinal data is however characterized by interdependencies such as multiple observations belonging to the same subject. This calls for the use of more complex statistical frameworks because simpler models (e.g., pre-post comparisons after aggregation over measurement occasions) cannot unveil more complex temporal dynamics which is actually a strength of intensive longitudinal studies. Simulation-based study planning allows researchers to adequately answer the substantive research question of app studies such as the one by Roepke et al. (2015) despite the high data complexity.

Among other frameworks such as linear growth models and cross-lagged panel models, (G)LMMs allow for adequately representing the data characteristics that come along with app-based psychological data collection. As Formulas (1) and (2) show, they can also be used to easily implement additional sources of data interdependence. In addition, they are able to account for data missingness, too (e.g., Hedeker & Gibbons, 2006). Given the wide applicability of (G)LMMs to simulate psychological app data, we will focus on them in the following.

The outcome of a single observation in *Who Knows*, Y_{ijk}, is the result of perceiver i judging target j on a single item k. Because i, j, and k contribute to multiple outcomes, observations are not independent. We account for this dependency by assuming a hierarchical data structure where each lower-level observation belongs to one higher-level combination of perceiver, target, and item, respectively. Since there can be an observation for each combination of higher-level units, the corresponding model is a *crossed random effects* model. In addition, we considered the respective combination of target j and item k to contribute to the outcome as a random interaction effect.

In *Who Knows*, the outcome of i judging j in terms of k (Y_{ijk}) can either be correct or incorrect and, thus, is a naturally dichotomous variable. As a result, the mathematical relationship between predictors and outcome is not linear. Rather, the effects of perceiver i, target j, item k, and the combination of j and k contribute to the probability of a correct outcome $P(Y_{ijk} = 1)$.

Binomial GLMMs (with a logit-link function) offer the possibility to model the non-linear relation between predictors and outcome within a hierarchical data structure. For this purpose, the linear combination of the random effects is inserted into the logistic formula:

$$P(Y_{ijk} = 1) = \frac{\exp(c_{000} + u_i^P + u_j^T + u_k^I + u_{jk}^{T\times I})}{1 + \exp(c_{000} + u_i^P + u_j^T + u_k^I + u_{jk}^{T\times I})}$$

u_i^P is the perceiver effect of *i*, u_j^T the target effect of *j*, and u_k^I the item effect of *k*. $u_{jk}^{T\times I}$ is the interaction effect of target *j* and item *k*. All these effects depict deviations from the mean intercept c_{000} that determines the average probability of a correct response. Because there are no predictors in our model, the resulting equation is a *random intercept-only model*. To obtain a dichotomous value of either 0 or 1 as the simulated outcome for one observation, we randomly drew a single value from a Bernoulli distribution with *p* equal to $P(Y_{ijk} = 1)$, thereby inducing unpredicted error into the outcome.

Choosing the parameters of the population model

After choosing an appropriate population model, it is necessary to consider possible choices for model parameters (e.g., effect sizes or regression coefficients). It is crucial to choose realistic values for the parameters and simulation conditions, respectively, because the validity of the simulation depends on it. Ideally, values can be inspired by existing literature or pilot studies having conducted similar analyses. In general, we recommend investigating a wide range of plausible values when uncertainty regarding a specific parameter is large. Exploring multiple plausible scenarios allows researchers to investigate the influence of the uncertain parameter value and thereby ensures that conclusions from the simulation do not depend too much on specific choices.

For variables related to the study design, such as study groups or time intervals between measurements, plausible values for the simulation should be derived from the study plan. Simulated values for time-variant and time-invariant predictors, on the other hand, must be drawn randomly from distributions. For each of these, plausible parameters must be set and assumptions on their distributions must be made prior to simulation to represent the distribution of the respective variable in the population. Time-invariant predictors or random effects in LMMs can be simulated from a single multivariate distribution, for instance, a normal distribution with mean vector μ and variance-covariance predictor matrix Σ. The assumption of normal distribution can also be violated to investigate the consequences of non-normality for the analyses (e.g., Auerswald & Moshagen, 2015). If the population model includes time-variant predictors, the multivariate distribution of the predictor variables must also describe auto-correlations of predictor variables and cross-lagged relations between predictor variables (Biesanz, 2012; Hertzog & Nesselroade, 2003).

Regardless of the statistical model, real outcome data can rarely be fully explained by the predictor variables which results in error terms adding "noise" to the data. Typically, the error is assumed to be independent from the other variables and, in linear models, a normal distribution centered at zero with error variance σ_ε^2 is often assumed:

$$\varepsilon \sim N(0, \sigma_\varepsilon^2) \quad (3)$$

Higher values for the error variance σ_ε^2 add more noise to the data. This simulates the case in which the predictors explain a smaller proportion of variance on the outcome, leading to a less precise estimation of the true model parameters in the later analysis. As the residuals are conceptualized as random variables, they can be drawn from any distribution, that is, the assumption of normally distributed residuals can be violated to observe its impact on the data analysis. It should be noted that a realistic choice for the level of noise variance is crucial for the validity of the simulation, especially too optimistic choices must be avoided. In the absence of similar previous studies, we recommend to use typical effect sizes of the respective field as a guideline what constitutes small, medium, and large effects. In the realm of psychological research, Cohen (1988) considered Pearson *r* values of 0.10, 0.30, and 0.50 to correspond to small, medium, and large effects, respectively. Median effect sizes, however, also depend on the sub-discipline (Schäfer & Schwarz, 2019). In personality psychology, for example, lower thresholds of 0.10, 0.20, and 0.30 are considered adequate (Gignac & Szodorai, 2016). We encourage researchers to take their time and thoughtfully select effect sizes that align as specifically as possible with the nuances and complexities of the planned study.

Our data generating model only contained the overall intercept c_{000} and four random effects u_i^P, u_j^T, u_k^I, and $u_{jk}^{T\times I}$. c_{000} determines the mean of the probability of a correct outcome $P(Y_{ijk} = 1)$, while the random effects represent whether specific perceivers, targets, and items, respectively, are judged more or less correctly than this average. All random effects were drawn from independent normal distributions with a corresponding mean of zero and variance σ_u^2: $N(0, \sigma_u^2)$. Altogether, we needed to choose values for five different parameters: c_{000}, as well as the four random effect variances.

On average, people make somewhat accurate personality judgments even on a very limited informational basis, mostly by using knowledge about broad social categories (*stereotype accuracy*, Jussim et al., 2009, 2015; e.g., based on gender, Löckenhoff et al., 2014; or age, Chan et al., 2012). We translated this principle into an above-chance mean probability of a correct outcome by setting c_{000} to 0.7. As a result, the mean correct outcome probability was 67%. Further, the extent to which people make accurate judgments should strongly depend on the combination of judged characteristic and target. Based on this, we assigned the highest effect variance to the target item interaction effect with $\sigma_{T\times I}^2 = 0.36$. The variances of the other three random main effects followed in descending order of assumed importance to variance explanation with perceiver variance $\sigma_P^2 = 0.125$, target variance $\sigma_T^2 = 0.04$, and item variance $\sigma_I^2 = 0.01$. The resulting random effect distributions are provided in Figure 1.

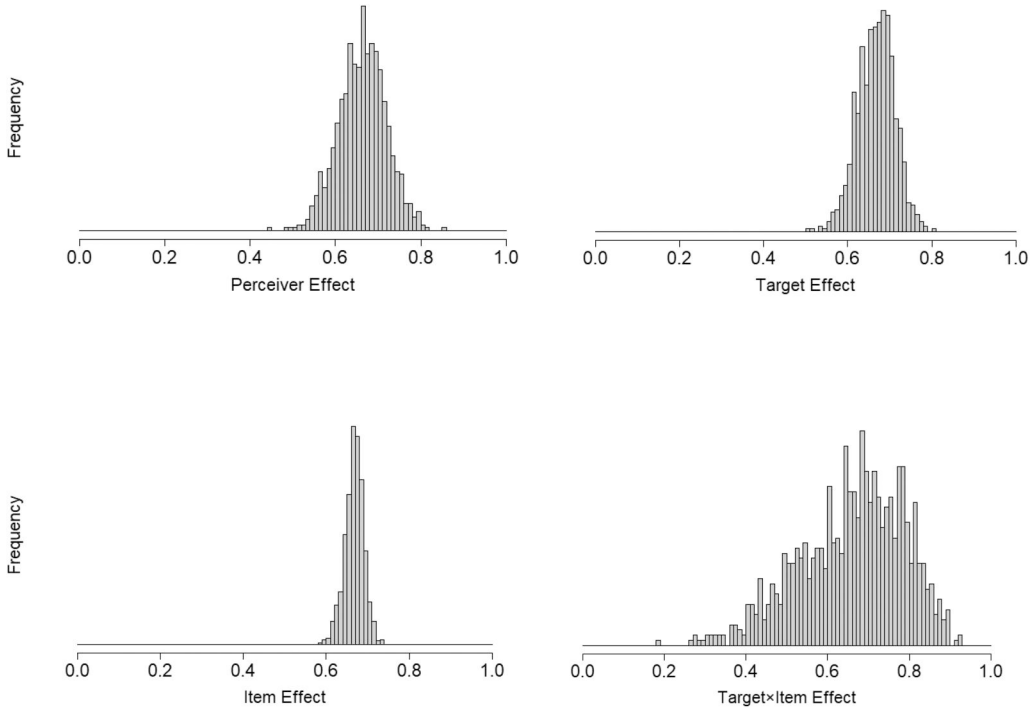

Figure 1. Random effect distributions used for simulating *Who Knows data*.
Note. The random effects were transformed into probability scale for the sake of interpretability. Following the GLMM population model, the original (normally distributed) random effects were generated on a logit-scale.

App usage scenarios

After implementing the steps from Section 2.1 in statistical software such as *R*, one can generate complete sample data for arbitrary sample sizes and parameter choices based on the defined population model and variables related to the study design. Another important consideration for the simulation is the choice of potential sample sizes. This is an especially difficult consideration for data collected with apps because both the number of observations and the pattern and degree of missingness in the data typically depend on user behavior which cannot be pre-planned. We therefore recommend investigating multiple usage scenarios representing rather pessimistic as well as rather optimistic expectations.

What are the sampling scheme and design factors?

To maximize control over missingness patterns, one should first create a complete dataset from the population and then randomly delete a number of values based on the anticipated missingness mechanism. The number of observations in the complete data set depends on the number of participants, study conditions, measurement occasions, items, and possibly additional design factors.

The amount of data required to answer one's research question is usually of great interest and a reason for conducting traditional power analyses. With this in mind, we recommend alternating the sample sizes or the numbers of items (or any other variable that affects the length of the final data set) between simulation conditions to gain insight into the required data quantity.

For *Who Knows* data generation, our goal was to first simulate the theoretically complete data set where every combination of perceiver, target, and item occurs exactly once. The result is a long-format data set where the total number of observations is the product of the numbers of perceivers (N_P), targets (N_T), and items (N_I), $N_P * N_T * N_I$.

Values for N_T and N_I were given by our app study design and set constant in each condition, to $N_T = 50$ and $N_I = 820$. The number of target $\times$ item combinations is the result of $N_T * N_I = 41{,}000$. However, because participation in our app study is entirely voluntary, the number of perceivers that engage in data collection could hardly be predicted. Since this is a crucial value for deriving strategies for advertising and recruiting, we were interested in how different N_P would affect the data and our simulation results. Therefore, we decided to simulate three different conditions for varying N_P with values of 300, 1000, and 5000. The lengths of the resulting long-format data sets of the three N_P conditions were 12.3 million, 41 million, and 205 million datapoints, respectively.

Planned missingness: Does every participant go through every study condition?

Planned missingness refers to missingness that is due to study design, where certain values are expected to be missing *a priori*. An example of this would be presenting only a subset of the total item pool to each participant, thereby reducing participant burden, and collecting data for a greater variety of items. The degree of planned missingness can also be considered a variable

design factor that can be varied between simulation conditions because it directly impacts the data quantity.

In the *Who Knows* app, targets' self-reports are available for only 60 out of the total 820 items for each target which results in data combinations for the rest of the items being planned as missing. We accounted for this by deleting observations for a random subset of 760 items for each target from the total data set created in the previous step.

Unplanned missingness: How frequently do participants use the app?

Another problem with remote data collection via an app is the lack of control over participation, which leads to high levels of dropouts (Meyerowitz-Katz et al., 2020; Torous et al., 2020). The resulting unplanned missingness is the main reason why we recommend simulating app data, because it creates difficulties in determining necessary sample sizes using traditional power analyses. Important considerations concern (1) the missingness mechanism, that is, whether data are missing (completely) at random or not at random, (2) how user behavior affects missingness, and (3) what pessimistic and optimistic scenarios could look like.

The nature of the missingness mechanism is defined by whether there are variables that explain the missingness of certain values. We refer to Schafer and Graham (2002) who developed a conceptualization that includes differentiating data that is *missing completely at random* (MCAR), *missing at random* (MAR), or *not at random* (MNAR). Missing data are MCAR if the probability of a missing value is not related to either another measured variable or the value itself. Data are MAR when the probability of a missing value is related to some other measured variable but not to the value itself. For data that is MNAR, the probability of a missing value is related to the value itself. This would be the case if patients with poor therapy outcomes were systematically missing at the last measurement occasion in the above-mentioned therapy study (Roepke et al., 2015).

Insights on reasons for dropout in remote studies and resulting missingness can be found in the literature (e.g., Brüggen & Dholakia, 2010; Nestler et al., 2015). If one concludes that the application of interest produces data that is not MCAR, the corresponding mechanism can be programmed into missingness introduction, i.e., data deletion, in the next step.[1]

[1]Systematic missingness can, for instance, be introduced to the simulation from a second GLMM with the missingness status (0 = not missing, 1 = missing) as dependent variable and whatever predictors seem suitable. The generated dichotomous random variable can then be used to filter out the missing values from the complete data set.

In accordance with the specified missingness mechanism, the user behavior and participation scheme contribute to missingness in the data. General attrition to app studies (see, e.g., Meyerowitz-Katz et al., 2020; Torous et al., 2020), enjoyable app experience, sample characteristics, and variance of dropout between participants are additional factors to consider in user behavior. If one's study collects time-series data, one needs to consider whether re-participation after absence at a given timepoint should be possible, and then establish appropriate missingness distributions for each timepoint.

Since all the considerations above are educated guesses at most, we recommend simulating optimistic and pessimistic missingness scenarios in separate conditions to evaluate the effect of missingness on the data. This applies to both the extend of missingness and the missingness mechanisms. Optimistic scenarios would then include cases with little to no dropout and/or data that is MCAR or at least MAR. More pessimistic scenarios would resemble cases in which many people drop out and the missingness mechanism is related to the missing value itself, resulting in data that is MNAR.

The most complete data the app can possibly produce—that is, with no dropouts, whatsoever– can serve as a reference condition. For an easy implementation of missingness into the data set, the *R* package *missMethods* (Rockel, 2022) can be used.

For our simulation, we assumed that no variable could explain the distribution of missingness on the outcome. As a result, data was MCAR and we were able to simulate missingness due to dropout by randomly deleting observations for the given participant perceivers.

Since participation in *Who Knows* data generation is entirely voluntary, the individual user behavior relates strongly to the degree of missingness. Consequently, we assumed that the degree of missing data varies between perceivers because some contributed more to data collection than others. To obtain the number of observations which needs to be set missing for each individual perceiver, we drew their individual number of games they *did* complete (their GPP; *games per perceiver*) from a skew-normal distribution truncated at the value 1 (perceivers need to play at least one game to be present in the data). For each perceiver, data for games that would exceed their GPP was deleted from the data set with planned missingness created in an earlier step.

We wanted to simulate three missingness scenarios that ranged from rather optimistic to pessimistic which we labeled *high, average,* and *low* participation, indicating the respective number of games in the condition. To do so, we specified three different parameters sets for the distributions from which individual GPPs were sampled. The distribution parameters were ξ = [20, 5, 1], ω = [50, 30, 10], and α = [10, 10, 10]. For 100 of the resulting distributions the average mean GPPs ($\overline{\text{GPPs}}$) were [59.45, 29.07, 4.98], the average median GPPs were [53, 24, 4], and the average skewnesses were [0.84, 0.94, 1.05]. For an outline of the resulting distributions, see Figure 2.

Although highly unrealistic, we also simulated the case where every participant completed every possible trial, resulting in 600 games for each participant and the most complete data possible, the *ideal* GPP condition. The results of this condition were planned to serve as a benchmark for the other three.

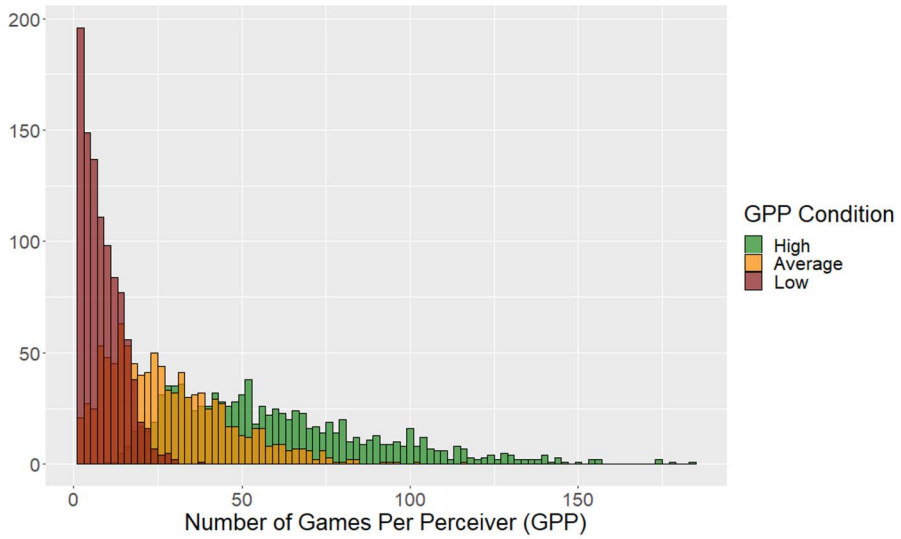

Figure 2. Skew-normal distributions for the number of games per perceiver (GPP) for three different GPP conditions in the simulation of who knows data.

Statistical analysis

Upon having generated a complete data set with the underlying model and deleting missing values, statistical analyses can be performed and the results for the given simulation parameters can be evaluated. For this purpose, two choices must be made: First, an appropriate framework for analyzing the given data structure and for answering the substantive research question must be selected. Second, adequate outcome metrics need to be chosen to decide on the study's scientific utility based on the analysis results.

What are the applied statistical frameworks?

Each simulated data set should be analyzed with an appropriate statistical method—mimicking the analyses that could be conducted with the real data set. Choices for an adequate framework consider its capacity to answer the research question and might consider robustness to missingness, computation time, or simplicity. It is also possible to conduct analysis using different techniques to weigh them up against each other and gain insight into which framework offers most advantages to data analysis under which data prerequisites. For instance, depending on the research questions, data sets that could be analyzed with (G)LMMs could also be analyzed using a two-step GLM approach (i.e., a GLM is fit for each level-2 unit and results are aggregated across all level-2 units). Using a GLM with cluster-robust standard errors, or generalized estimation equations. These data-analytic approaches can be directly compared in the simulation.

As we explained earlier, a good choice to assess whether the true model parameters of the simulation can be recovered reliably is to use same statistical framework for analysis as has been used for the simulation before. Apart from that, the influence of different analytical choices can be compared directly. For instance, one might want to compare different options to handle missingness, such as full information maximum likelihood (Schafer & Graham, 2002; Wickham & Giordano, 2022), multiple imputation (Graham et al., 1996), or Markov Chain Monte Carlo algorithm (Wickham & Giordano, 2022) over one that is not (e.g., last observation carried forward, Zhu, 2014; listwise and pairwise deletion, Newman, 2003). Importantly, each simulated data set should be analyzed with the full intended analysis pipeline including preprocessing steps.

For data analysis, we were interested in whether the true individual random effects u_i^P, u_j^T, u_k^I, and $u_{jk}^{T\times I}$ could be reliably recovered. For this matter, we analyzed the data with the same framework we used for data generation: the binomial GLMM.

However, because of the long computation time that was needed for binomial GLMM estimation, we also estimated the individual random effects by aggregating the associated outcomes into mean scores. Compared to the GLMM estimation, we expected this *aggregated mean approach* to provide advantages in computation time but also to yield less reliable estimates. As such, pitting both of these approaches against each other when analyzing the simulated data would clarify which approach would be preferable in which data scenario.

What are the outcome metrics of interest and what are criteria for scientific utility?

It is important to define criteria for acceptable and unacceptable data quality before running the simulation

study. For this matter, the goal is to (1) find appropriate outcome metrics for the research question of the simulation and (2) formulate adequate thresholds for it. Depending on the research question, different outcome metrics can be computed. For example, if the simulation is supposed to give insight on estimation accuracy of the true effects, the outcome metric of choice might be estimation bias (i.e., the difference between the estimated and the true value), estimation error (i.e., the empirical standard error) or coverage of confidence intervals (i.e., the proportion of samples in which the CIs contained the true value). If one is interested in the statistical power of an effect, the corresponding outcome metric would be the proportion of samples in which the null hypothesis for that effect is rejected (when there actually is an effect).

Ideally, for each metric of interest, thresholds should be defined to provide a final evaluation of the app's scientific utility under different simulation conditions. This allows to declare certain conditions as insufficient to answer one's research question. For instance, a relative estimation bias of $\leq |10\%|$ for a given effect could be considered sufficient. Regarding statistical power, a threshold of 80% is often considered sufficient for psychological studies (Cohen, 1992) and is commonly used as a benchmark in power simulation studies (e.g., Brysbaert & Stevens, 2018). While guidelines for suitable thresholds can often be found in methodological papers, there is no universal recommendation, as the choice of thresholds heavily relies on the specific research area and its implications. For example, in scenarios where an unreliable estimate or a falsely negative test result carries significant consequences (e.g., when screening for suicidal risk), aiming for higher diagnostic precision or statistical power than generally recommended is warranted. However, in other instances, the costs of pursuing high precision or power may not be justified. Valuable sources for determining appropriate thresholds include insights from previous studies on similar topics and recommendations from professional organizations in the field.

The main goal of the *Who Knows* application is to differentiate perceivers based on their ability to judge targets in terms of everyday life characteristics. In follow-up studies, this variance of the perceiver effect might be explained by other third variables, such as the Big Five personality traits. Another possible application might be to make decisions for individuals based on their score in *Who Knows*, like finding the "best judge" amongst a group of colleagues. For both of these matters, the application needs to accurately recover the true rank-order of perceiver effects. Consequently, the metric that we were most interested in was the reliability of perceiver effects which we computed as the Pearson correlation between the simulated true perceiver effects and the perceiver effects estimated in our analyses.

Following Nunnally's (1978) recommendations, we established different reliability thresholds to ultimately evaluate the simulation. We used a threshold of $rel = 0.80$ to identify simulation conditions that are suitable to answer group level research questions, such as the relation between individuals' ability to judge others in *Who Knows* and their agreeableness. Further, we considered conditions with reliabilities of $rel \geq 0.90$ appropriate for guiding individual-level decisions based on *Who Knows* scores. These thresholds were meant to guide our overall evaluation of the scientific utility of data collection with the *Who Knows* app across various scenarios.

Procedure/pseudo code

The present considerations regarding the simulation preparation must be implemented into software code to generate and analyze data. Independent of the particular software in use, the logic of the simulation procedure can be summarized by the following pseudo code:

```
SET condition values
SET number of samples
SET population parameter values
FOR each simulation condition
  FOR each simulated sample
    COMPUTE complete data set with population model and parameters
    APPLY planned missingness to dataset (optional)
    APPLY unplanned missingness to dataset (optional)
    APPLY preprocessing steps to dataset (optional)
    APPLY statistical model(s) to dataset
    COMPUTE relevant metrics
    SAVE raw dataset and analysis results
  END FOR
END FOR
FOR each simulation condition
  SUMMARIZE metrics
END FOR
```

For Who Knows, we used the statistics software R, version 4.2.1 (R Core Team, 2022). The simulation scripts can be retrieved from tinyurl.com/appstudysimulation.

The simulation results showed that perceiver reliability is merely a function of perceiver missingness or, in other words, the GPP. Because perceivers using *Who Knows* will vary greatly in their individual GPP the question arises for whom there will be sufficient data to achieve reliabilities of 0.80 or 0.90 in the data. By subdividing the perceivers in our simulated data into groups of specific GPP ranges, we were able to calculate and plot the perceiver reliability as a function of GPP (see Figure 3). This revealed that perceivers need to play 30 to 40 games (and rate 150 to 200 items) to yield reliabilities of 0.80 or above. For perceiver reliabilities above 0.90, perceivers need to play 70 to 80 games.

These findings have direct implications for data collection with *Who Knows*. They show that it might be reasonable to exclude data from perceivers with a GPP less than 30 from further analysis because their data is not reliable enough to draw scientific conclusions. Another implication is that more emphasis should be placed on incentivizing existing perceivers in the app to play a minimum of 30 games than on acquiring new perceivers. This insight found its way into the design of the real *Who Knows* app by allowing perceivers to unlock the available feedback about their performance only after reaching certain experience levels.

Our *Who Knows* simulation aimed for recovering the true rank-order of random effects. More specifically, the goal was to identify simulation conditions that could recover the rank-order of perceiver effects

sufficiently well, with perceiver reliabilities of at least 0.80 for group-level research on perceiver effects and reliabilities of at least 0.90 for individual-level research. Table 2 gives an overview of the reliabilities as a function of analysis approach, the number of perceivers N_P, and the GPP condition for each different random effect. The ideal condition resembles the highly unlikely scenario of every perceiver having engaged in every possible trial and generated every possible combination of perceiver, target, and item. We included it into the simulation and analysis as reference for the results yielded for other simulation conditions and to understand the impact of unplanned missingness on the results.

As for perceivers, every reliability in the GPP conditions with ideal and high $\overline{GPP}$, independent of statistical approach or N_P, exceeded .80. Reliabilities of perceiver effects exceeded 0.90 only in the ideal GPP conditions. Therefore, if our simulation parameters and assumptions were realistic, *Who Knows* can only be used to investigate research questions that relate to group-level (rather than individual-level) perceiver effects, as we do not expect the ideal GPP conditions to represent realistic scenarios. Consequently, if individual estimates are of interest, perceivers will need to be externally incentivized to play a larger number of games as most of them would play just for fun.

Results furthermore showed slight advantages of the GLMM approach over the mean-based approach in terms of recovering true random effects. This effect grew larger for poor data conditions (e.g., those with $\overline{GPP} = 5$ and $N_P = 300$) with a high degree of missingness. On the other hand, GLMM estimations was sometimes a matter of hours for a single data set, especially for good data conditions with many observations (e.g., those with $\overline{GPP} = 61$ and $N_P = 5000$). We conclude that for analyses of *Who Knows* data GLMM estimation is to be preferred for poor data conditions, whereas for good data conditions the mean-based approach represents the better choice, as it yields similarly accurate results in only a fraction of computation time.

Concluding remarks

Our approach has several advantages over traditional approaches performed to identify data prerequisites before collection, such as power analyses. Foremost, given enough computational resources, it can incorporate arbitrary degrees of data complexity regarding interdependencies and missingness, as well as non-linear relations or complex dynamics in time series. Our simulation of *Who Knows* data serves as a good example for the complex intertwinement of different data characteristics that can be generated, as it implemented multilevel structures, different degrees of missingness, and a non-linear logistic response function between predictors and outcome. Conventional power analyses, on the other hand, lack the flexibility to be used for more complex applications (Brysbaert & Stevens, 2018; Green & MacLeod, 2016). Moreover, they cannot investigate the effect of missingness on the data, whereas Monta Carlo simulations have been explicitly recommended as a viable solution to this (Davey & Savla, 2010; Wu, 2004). Simulations also provide the opportunity to observe the effects of certain parametric or distributional assumptions on the performance of analysis approaches, which is typically not possible with analytic power analyses.

While many model parameters need to be anticipated to simulate data and the generalizability of the simulation results is limited to these parameters, power analyses are subject to the same limitations, too. In fact, the sparse basis on which parameters for power analyses are often guessed has been criticized before (Gelman & Carlin, 2014). By systematically altering certain simulation parameters between different conditions, our approach can identify those parameters that are critical to the scientific utility of one's app study and derive direct measures for

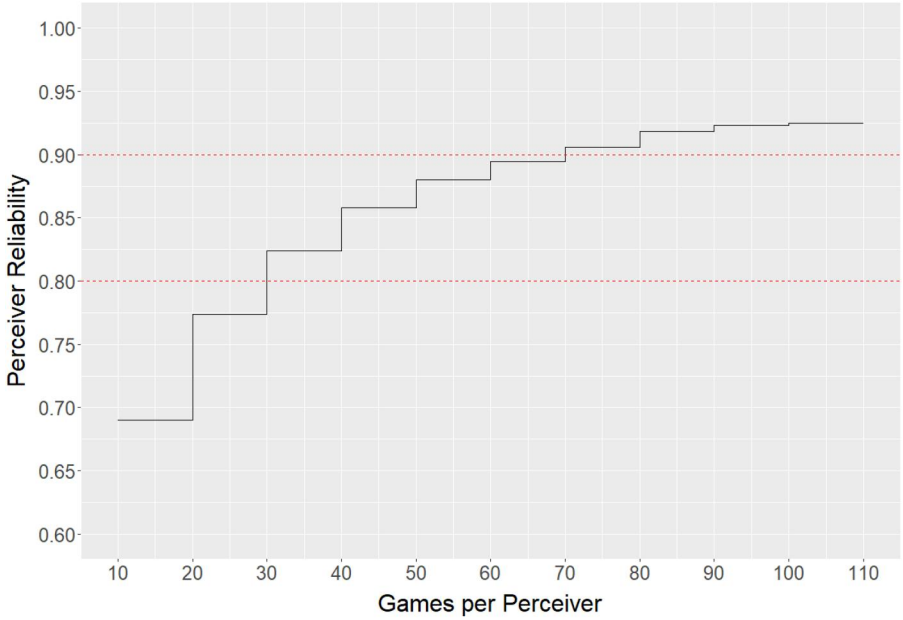

Figure 3. Perceiver reliabilities as a function of the number of games per perceiver (GPP).

Table 2. Reliabilities of the four random effects as a function of analysis approach, N_P, and GPP condition.

			Random effects			
Approach	N_P	GPP condition	Perceivers	Targets	Items	Target × Item
GLMM	5000	Ideal	**.99***	**.93***	.30	**.98***
		High	.87*	**.93***	.30	**.97***
		Average	.77	**.92***	.30	**.95***
		Low	.48	**.91***	.27	.85*
	1000	Ideal	**.99***	**.93***	.30	**.97***
		High	.87*	**.92***	.29	**.92***
		Average	.77	**.91***	.27	.87*
		Low	.48	.84*	.19	.61
	300	Ideal	**.99***	**.93***	.30	**.96***
		High	.87*	**.90***	.26	.81*
		Average	.77	.87*	.22	.70
		Low	.47	.69	.12	.40
Mean-based	5000	Ideal	**.99***	**.93***	.26	**.93***
		High	.85*	**.93***	.25	**.92***
		Average	.71	**.92***	.25	**.91***
		Low	.40	**.90***	.23	.82*
	1000	Ideal	**.99***	**.93***	.25	**.93***
		High	.86*	**.92***	.24	.88*
		Average	.71	**.91***	.23	.84*
		Low	.41	.83*	.16	.58
	300	Ideal	**.99***	**.92***	.25	**.91***
		High	.86*	**.90***	.22	.79
		Average	.71	.87*	.19	.68
		Low	.40	.68	.10	.36

Note. * > 0.80; reliabilities > 0.90 are presented in bold font.

practice. For example, by simulating different degrees of data missingness in terms of different GPP distributions, we inferred the importance of the GPP in recovering the rank-order of perceiver effects. We incorporated this finding into real data generation by increasing the focus on individual perceiver feedback as an incentivization, especially for playing the first 30 games in *Who Knows*. In either case, we recommend cross-checking the simulation result patterns for plausibility with findings from the literature. For instance, the relationship between the GPP and the reliability of perceiver effects is supported by the psychometric literature, as it mimics the Spearman-Brown relation (Brown, 1910; Spearman, 1910). Moreover, there are aspects of real data generation that should be reflected in the simulation, but over which one has no control (e.g., the dropout rate). Therefore, to assess how realistic the simulated scenarios were, they should be compared with the real data as soon as these become available. In our case, a comparison with real *Who Knows* data showed that the true distribution of GPPs has a mean GPP of 23.60, a median GPP of 18 and a skewness of 2.09 and is thus similar (except for the skewness) to the average GPP condition defined by us.

There are two prerequisites of implementing the simulation approach which might seem like obstacles that are not present in traditional power analyses: the choice of an appropriate mathematical framework for data generation and its translation into code. We argue, however, that meeting these prerequisites hardly exceeds the effort of a power analysis: Once an appropriate mathematical framework for the analysis is found, the same framework can be used for data generation, too. Finding a suitable framework for analysis, however, is not a prerequisite exclusive to the simulation approach, but is necessary to accurately represent true data relations and derive correct conclusions from any scientific study. Therefore, the mathematical complexity of the simulation study does typically not exceed the complexity of the analysis itself. In fact, running a simulation study before data collection may help researchers to explicitly plan statistical analyses beforehand which is also necessary for preregistrations. The subsequent translation of the statistical framework into code is moreover facilitated by packages that are available for the prominent programming languages, such as *R*. This way, the simulation approach is accessible to researchers with little programming knowledge, too.

Conclusion

The present work highlighted the usefulness of conducting a Monte Carlo simulation to evaluate the scientific utility of an app-based psychological study prior to data collection. In addition, we provided a set of questions that can guide researchers in conducting their own simulation. App-based data collection promises great benefits in psychological research but the complex data structures resulting from them can make it hard to know ahead of time whether a particular study will pay off scientifically. The present work showcased the usefulness of Monte Carlo simulations to forecast the scientific utility of app studies by flexibly accounting for the interdependence and missingness of observations which are common in these data environments. Although there may be edge cases for which our propositions are incomplete, they can guide the general thought processes and steps that researchers would follow in implementing a particular simulation. With the present tutorial, researchers now have checklist and a set of recommendations at hand with which they can evaluate in advance the scientific utility of planned app study studies such that they can take full advantage of the promises of mobile data collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not funded.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

Open Scholarship

This article has earned the Center for Open Science badges for Open Materials. The materials are openly accessible at https://OSF.IO/WSKGH.

References

Alkhaldi, G., Hamilton, F. L., Lau, R., Webster, R., Michie, S., & Murray, E. (2016). The effectiveness of prompts to promote engagement with digital interventions: A systematic review. *Journal of Medical Internet Research*, *18*(1), e6. https://doi.org/10.2196/jmir.4790

Auerswald, M., & Moshagen, M. (2015). Generating correlated, non-normally distributed data using a non-linear structural model. *Psychometrika*, *80*(4), 920–937. https://doi.org/10.1007/s11336-015-9468-7

Ben-Zeev, D., Brenner, C. J., Begale, M., Duffecy, J., Mohr, D. C., & Mueser, K. T. (2014). Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. *Schizophrenia Bulletin*, *40*(6), 1244–1253. https://doi.org/10.1093/schbul/sbu033

Biesanz, J. (2012). Autoregressive longitudinal models. *Handbook of Structural Equation Modeling*, 459–471. https://psycnet.apa.org/record/2012-16551-027

Brouillette, R. M., Foil, H., Fontenot, S., Correro, A., Allen, R., Martin, C. K., Bruce-Keller, A. J., & Keller, J. N. (2013). Feasibility, reliability, and validity of a smartphone based application for the assessment of cognitive function in the elderly. *PloS One*, *8*(6), e65925. https://doi.org/10.1371/journal.pone.0065925

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904-1920*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Brüggen, E., & Dholakia, U. M. (2010). Determinants of participation and response effort in web panel surveys. *Journal of Interactive Marketing*, *24*(3), 239–250. https://doi.org/10.1016/j.intmar.2010.04.004

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 9. https://doi.org/10.5334/joc.10

Chan, W., Mccrae, R. R., De Fruyt, F., Jussim, L., Löckenhoff, C. E., De Bolle, M., Costa, P. T., Sutin, A. R., Realo, A., Allik, J., Nakazato, K., Shimonaka, Y., Hřebíčková, M., Graf, S., Yik, M., Brunner-Sciarra, M., De Figueroa, N. L., Schmidt, V., Ahn, C.-K., … Terracciano, A. (2012). Stereotypes of age differences in personality traits: Universal and accurate? *Journal of Personality and Social Psychology*, *103*(6), 1050–1066. https://doi.org/10.1037/a0029712

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.). L. Erlbaum Associates.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Davey, A., & Savla, J. (2010). *Statistical power analysis with missing data: A structural equation modeling approach*. Routledge.

Fischer, F., & Kleen, S. (2021). Possibilities, problems, and perspectives of data collection by mobile apps in longitudinal epidemiological studies: Scoping review. *Journal of Medical Internet Research*, *23*(1), e17691. https://doi.org/10.2196/17691

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*(3), 177–182. https://doi.org/10.1177/0963721412445309

Gelman, A., & Carlin, J. B. (2014). Beyond power calculations. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*(2), 197–218. https://doi.org/10.1207/s15327906mbr3102_3

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using

smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Hedeker, D. R., & Gibbons, R. D. (2006). *Longitudinal data analysis. Wiley & Sons Ltd.*

Hertzog, C., & Nesselroade, J. R. (2003). Assessing psychological change in adulthood: An overview of methodological issues. *Psychology and Aging*, *18*(4), 639–657. https://doi.org/10.1037/0882-7974.18.4.639

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 199–227). Psychology Press.

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in) accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science*, *24*(6), 490–497. https://doi.org/10.1177/0963721415605257

Löckenhoff, C. E., Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., De Bolle, M., Costa, P. T., Jr, Sutin, A. R., Realo, A., Allik, J., Nakazato, K., Shimonaka, Y., Hřebíčková, M., Graf, S., Yik, M., Ficková, E., Brunner-Sciarra, M., Leibovich de Figueora, N., Schmidt, V., Ahn, C-k., Ahn, H-n., … Terracciano, A. (2014). Gender stereotypes of personality. *Journal of Cross-Cultural Psychology*, *45*(5), 675–694. https://doi.org/10.1177/0022022113520075

Meyerowitz-Katz, G., Ravi, S., Arnolda, L., Feng, X., Maberly, G., & Astell-Burt, T. (2020). Rates of attrition and dropout in app-based interventions for chronic disease: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *22*(9), e20283. https://doi.org/10.2196/20283

Miner, A., Kuhn, E., Hoffman, J. E., Owen, J. E., Ruzek, J. I., & Taylor, C. B. (2016). Feasibility, acceptability, and potential efficacy of the PTSD Coach app: A pilot randomized controlled trial with community trauma survivors. *Psychological Trauma: Theory, Research, Practice and Policy*, *8*(3), 384–392. https://doi.org/10.1037/tra0000092

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Mulder, J. D. (2022). Power analysis for the random intercept cross-lagged panel model using the powRICLPM R-package. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(4), 645–658. https://doi.org/10.1080/10705511.2022.2122467

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide*. (8th ed.). Muthén & Muthén.

Nestler, S., Thielsch, M., Vasilev, E., & Back, M. D. (2015). Will they stay or will they go? Personality predictors of dropout in an online study. *International Journal of Internet Science*, *10*(1), 37–48.

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, *6*(3), 328–362. https://doi.org/10.1177/1094428103254673

Nunnally, J. C. (1978). An overview of psychological measurement. *Clinical Diagnosis of Mental Disorders*, 97–146. https://doi.org/10.1007/978-1-4684-2490-4_4

Pavliscsak, H., Little, J. R., Poropatich, R. K., McVeigh, F. L., Tong, J., Tillman, J. S., Smith, C. H., & Fonda, S. J. (2016). Assessment of patient engagement with a mobile application among service members in transition. *Journal of the American Medical Informatics Association: JAMIA*, *23*(1), 110–118. https://doi.org/10.1093/jamia/ocv121

Pfammatter, A. F., Mitsos, A., Wang, S., Hood, S. H., & Spring, B. (2017). Evaluating and improving recruitment and retention in an mHealth clinical trial: An example of iterating methods during a trial. *mHealth*, *3*, 49–49. https://doi.org/10.21037/mhealth.2017.09.02

Pornprasertmanit, S., Miller, P., Schoemann, A., Jorgensen, T. D. (2021). *simsem: SIMulated structural equation modeling. R package version 0.5*-16. https://CRAN.R-project.org/package=simsem.

Rau, R., Grosz, M. P., & Back, M. D. (2023). A large-scale, gamified online assessment of first impressions: The Who Knows project. https://doi.org/10.31234/osf.io/gb4av

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rockel, T. (2022). *missMethods: Methods for missing data. R package version 0.4*.0, https://CRAN.R-project.org/package=missMethods.

Roepke, A. M., Jaffee, S. R., Riffle, O. M., McGonigal, J., Broome, R., & Maxwell, B. (2015). Randomized controlled trial of SuperBetter, a smartphone-based/internet-based self-help tool to reduce depressive symptoms. *Games for Health Journal*, *4*(3), 235–246. https://doi.org/10.1089/g4h.2014.0046

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 813. https://doi.org/10.3389/fpsyg.2019.00813

Smith, A., de Salas, K., Lewis, I., & Schüz, B. (2017). Developing smartphone apps for behavioural studies: The AlcoRisk app case study. *Journal of Biomedical Informatics*, *72*, 108–119. https://doi.org/10.1016/j.jbi.2017.07.007

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Torous, J., Lipschitz, J., Ng, M., & Firth, J. (2020). Dropout rates in clinical trials of smartphone apps for depressive symptoms: A systematic review and meta-analysis. *Journal of Affective Disorders*, *263*, 413–419. https://doi.org/10.1016/j.jad.2019.11.167

Wenz, A., Jäckle, A., Burton, J., & Couper, M. P. (2022). The effects of personalized feedback on participation and reporting in mobile app data collection. *Social Science Computer Review*, *40*(1), 165–178. https://doi.org/10.1177/0894439320914261

Wickham, R. E., & Giordano, B. L. (2022). Implementing planned missingness in stimulus sampling designs: Strategies for optimizing statistical power and precision while limiting participant burden. *Journal of Experimental Social Psychology*, *101*, 104349. https://doi.org/10.1016/j.jesp.2022.104349

Wu, L. (2004). Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. *Journal of the American Statistical Association*, *99*(467), 700–709. https://doi.org/10.1198/016214504000001006

Yarkoni, T. (2022). The generalizability crisis. *The Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Zhu, X. (2014). Comparison of four methods for handing missing data in longitudinal data analysis through a simulation study. *Open Journal of Statistics*, *04*(11), 933–944. https://doi.org/10.4236/ojs.2014.411088

Zorluoglu, G., Kamasak, M. E., Tavacioglu, L., & Ozanar, P. O. (2015). A mobile application for cognitive screening of dementia. *Computer Methods and Programs in Biomedicine*, *118*(2), 252–262. https://doi.org/10.1016/j.cmpb.2014.11.004