¹ **Accuracy in Parameter Estimation and Simulation Approaches for Sample Size**

² **Planning with Multiple Stimuli**

³ Erin M. Buchanan[1] & Other Folks as Per Order on Doc[2]

⁴ [1] Harrisburg University of Science and Technology

⁵ [2] Other Instituions

⁶ **Author Note**

⁷

⁸   The authors made the following contributions. Erin M. Buchanan:

⁹ Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing,

¹⁰ Resources, Validation, Visualization, Project Administration, Formal Analysis; Other Folks

¹¹ as Per Order on Doc: Writing - Review & Editing, Data Curation, Resources.

¹²   Correspondence concerning this article should be addressed to Erin M. Buchanan,

¹³ 326 Market St., Harrisburg, PA, 17101. E-mail: ebuchanan@harrisburgu.edu

Abstract

14

15   The planning of the sample size for research studies often focuses on obtaining a significant

16   result given a specified level of power, significance, and an anticipated effect size. This

17   planning requires prior knowledge of the study design and a statistical analysis to calculate

18   the proposed sample size. However, there may not be one specific testable analysis from

19   which to derive power (Silberzahn et al., 2018) or a hypothesis to test for the project (e.g.,

20   creation of a stimuli database). Modern power and sample size planning suggestions

21   include accuracy in parameter estimation (AIPE, Kelley, 2007; Maxwell et al., 2008) and

22   simulation of proposed analyses (Chalmers & Adkins, 2020). These toolkits provide

23   flexibility in traditional power analyses that focus on the if-this, then-that approach, yet,

24   both AIPE and simulation require either a specific parameter (e.g., mean, effect size, etc.)

25   or statistical test for planning sample size. In this tutorial, we explore how AIPE and

26   simulation approaches can be combined to accommodate studies that may not have a

27   specific hypothesis test or wish to account for the potential of a multiverse of analyses.

28   Specifically, the examples focus on studies that adopt multiple items and suggest that

29   sample sizes can be planned to measure those items adequately and precisely, regardless of

30   statistical test. We demonstrate that pilot data can be used to determine a sample size

31   that represents well-measured data. This tutorial also provides multiple code vignettes

32   that researchers can adapt and apply to their own measures.

33   *Keywords:* accuracy in parameter estimation, power, sampling, simulation,

34   hypothesis testing

Accuracy in Parameter Estimation and Simulation Approaches for Sample Size Planning with Multiple Stimuli

Statistical power and power analyses are arguably one of the most important components in planning a research study (Cohen, 1990). However, if reviews of transparency and openness in research publications are any clue, the social sciences have failed to fully implement power analyses as part of their common efforts (Hardwicke et al., 2020, 2022). The replication "crisis" and credibility revolution have shown that published studies in psychology are underpowered (Open Science Collaboration, 2015; Vazire, 2018). Pre-registration of a study involves outlining the study and hypotheses before data collection begins (Chambers et al., 2014; Nosek & Lakens, 2014; Stewart et al., 2020), and details of a power analyses or other limitations on resources are often used to provide justification for the pre-registered sample quota. Given the combined issues of publish-or-perish and that most non-significant results do not result in published manuscripts, one may expect that power analysis would be especially critical for early career researchers (Rosenthal, 1979; Simmons et al., 2011). At best, an underpowered study provides limited insight (Halpern, 2002), and it can be difficult to know if a poorly implemented power analysis is better than no power analysis.

A recent review of power analyses found - across psychology journal articles - researchers did not provide enough information to understand their power analyses and often chose effect sizes that were unjustified (Beribisky, 2019). One solution to this power analysis problem is the plethora of tools made available for researchers to make power computations accessible to non-statisticians with the caveat that a solid education in power is necessary to use these tools. G*Power is one of the most popular free power software options (Erdfelder et al., 1996; Faul et al., 2007) that provides a simple point and click graphical user interface for power calculations (however, see Brysbaert, 2019). Web-based tools have also sprung up for overall and statistical test specific sample size planning

including https://powerandsamplesize.com and https://designingexperiments.com (Anderson et al., 2017). *R*-coding based packages, such as *pwr* (Champely et al., 2017), *faux* (DeBruine, 2021), and *SimDesign* (Chalmers & Adkins, 2020), can be used to examine power and sample size planning usually with simulation. Researchers must be careful using any toolkit, as errors can occur with the over-reliance on software (e.g., it should not be a substitute for critical thinking, Nuijten et al., 2016). When computing sample size estimates, it is important to remember that these values are estimations, not exact calculations guaranteed to produce a specific result (Batterham & Atkinson, 2005). For example, if any parameter estimated by the researcher was not found in the study (i.e., a smaller effect size than used for the power analysis), then the sample size estimate tied to that specific level of power may be incorrect.

Changes in publication practices and research design have also created a new wrinkle in providing a sample size plan for a research study. While statistics courses often suggest that a specific research design leads to a specific statistical test, meta-science work has shown that given the same data and hypothesis, researchers can come up with multiple ways to analyze the data Coretta et al. (2023). Therefore, a single power analysis only supports the specific analysis in which the researcher expects to test. Analyses may evolve during the research project or be subject to secondary analysis; thus, power and sample size estimation based on one analysis is potentially less useful than previously imagined. Further, research projects often have multiple testable hypotheses, but it is unclear which hypothesis or test should be used to estimate sample size with a power analysis. Last, research publications may not even have a specific, testable hypothesis, as some publications are intended to curate a large dataset for future reuse (i.e., stimuli database creation, Buchanan et al., 2019).

In light of these analytical (or lack thereof) concerns, we propose a new method that combines accuracy in parameter estimation Maxwell et al. (2008) and bootstrapped

simulation on pilot data (Rousselet et al., 2022). This method accounts for a potential lack

of hypothesis test (or simply not a good way to estimate an effect size of interest), and/or

an exploratory design with an unknown set of potential hypotheses and analytical choices.

Additionally, we consider the nature of cognitive research designs that use multiple items

to measure the phenomena of interest. For example, semantic priming is measured with

multiple paired stimuli (Meyer & Schvaneveldt, 1971), which traditionally was analyzed by

creating person or item-level averages for an ANOVA (Brysbaert & Stevens, 2018).

However, the use of multilevel models with random effects for the stimuli used in a study

have shown that we should be careful to assume that all items of a research study have the

same "effect", as there is often variability in their impact on the outcomes of the study.

Accuracy in Parameter Estimation

AIPE shifts the focus away from finding a significant p-value to finding a parameter

that is "accurately measured". For example, researchers may wish to detect a specific

correlation in a study, $r = .35$. They could then use AIPE to estimate the sample size

needed to find a "sufficiently narrow" confidence interval around that correlation.

Sufficiently narrow is often defined by the researcher using a minimum parameter size of

interest and confidence intervals. Therefore, they could decide that their 95% confidence

interval should be approximately between .20 and .50, and sufficiently narrow could be

defined as a width of .30 or .15 on each side. While confidence intervals are related to null

hypothesis significance testing (i.e., 95% confidence intervals that do not include zero

would indicate a significant difference from zero at $\alpha < .05$), AIPE procedures suggest how

we can define a sample size with a given width of confidence interval, regardless of whether

it includes zero.

Bootstrapping and Simulation

Bootstrapping involves using data obtained to simulate similar datasets by drawing

from the original data with replacement (Efron, 2000; Rousselet et al., 2022).

Bootstrapping is a form of data simulation that allows one to calculate parameter

estimates, confidence intervals, and more to simulate the potential population distribution, shape, and bias. Simulation is often paired with making up data for testing analyses and hypotheses based on proposed effect size or suggested population means. Generally, we would suggest starting with pilot data of a smaller sample size to understand the variability in potential items used to represent your phenomenon, especially if they are to be used in a larger study. However, given some background knowledge about the potential items, one could simulate example pilot data to use in. similar manner in our suggested procedure. Pilot or simulated data would be used to estimate the variability within items and select a "sufficiently narrow" window for item's confidence interval for AIPE (i.e., by selecting a specific standard error criterion, given the formula for confidence intervals). Bootstrapping would then be used to determine how many participants may be necessary to achieve a dataset wherein many items meet the required confidence interval.

Sequential Testing

Researchers could then use sequential testing to estimate their parameter of interest after each participant's data was collected to determine whether they have achieved their expected width of the confidence interval around that parameter. One would set a minimum sample size (i.e., based on known data collection ability) and use the confidence interval width as a stopping rule (i.e., stop data collection when the CI is narrow, as defined above). Next, researchers would use the estimated sample size associated with the simulation results of many items obtaining the stopping rule as a maximum sample size (i.e., we expect 90% of items to meet our stopping rule with 100 participants based on simulation). By defining each of these components, researchers could ensure a feasible minimum sample size, a way to stop data collection when goals have been met, and a maximum sample size rule to ensure an actual end to data collection. Therefore, we proposed that we should be able to leverage the ideas behind AIPE, paired with simulation and bootstrapping, to estimate the minimum and maximum proposed sample sizes and stopping rules for repeated measures studies with expected variability in parameter

141 estimates for items.

Proposed Method to Calculate Sample Size

143 Using these ideas, we suggest the following procedure to determine a sample size for
144 each item:

145 1) Use pilot data that closely resembles data you intend to collect. This dataset should
146 contain items that are identical or similar to those that will be collected in the study.
147 In this procedure, it is important to ensure that the data is representative of a larger
148 population of sampled items that you intend to assess. Generally, pilot data sample
149 sizes will be small, as the goal would be to determine how many items would be
150 necessary to reach a "stable" standard error.

151 2) Calculate the standard error (SE) in the subset of pilot data associated with each
152 item to create a cutoff score that defines when items are considered "accurately
153 measured". Next, the simulations below will explore what criterion should be used to
154 determine the cutoff score from the pilot data.

155 3) Sample, with replacement, from your pilot data using sample sizes starting at 20
156 participants and increase in small units up to a value that you consider the maximum
157 sample size. We will demonstrate example maximum sample sizes based on the data
158 simulation below; however, a practical maximum sample size may be determined by
159 time (e.g., one semester data collection) or resources (e.g., 200 participants worth of
160 funding). Although 20 participants would likely yield imprecise estimates, we suggest
161 this starting minimum for simulation purposes.

162 4) For each simulation, calculate the standard error for each item, and use these values
163 to determine the percentage of items that meet the cutoff score determined in Step 2.

164 5) Find the minimum sample size so that 80%, 85%, 90%, and 95% of the items meet

165   the cutoff score and can be considered accurately measured. We recommend these

166   scores to ensure that most items are accurately measured, in a similar vein to the

167   common power-criterion suggestions. Each researcher can determine which of these is

168   their minimum or maximum sample size (e.g., individuals can choose to use 80% as a

169   minimum and 90% as a maximum or use values from Step 3 based on resources).

170   6) Report these values, and designate a minimum sample size, the cutoff criterion, and

171   the maximum sample size. Each researcher should also report if they plan to use an

172   adaptive design, which would stop data collection after meeting the cutoff criterion

173   for each item.

174   These steps are summarized in Table 1. We will first demonstrate the ideas behind

175   the steps using open data (Balota et al., 2007; Brysbaert et al., 2014). This example will

176   reveal a few areas of needed exploration for the steps. Next, we portray simulations for the

177   proposed procedure and find solutions to streamline and improve the sample size

178   estimation procedure. Finally, we include additional resources for researchers to use to

179   implement the estimation procedure.

Example

181   In this section, we provide two examples of the suggested procedure. The first

182   example includes concreteness ratings from Brysbaert et al. (2014). Instructions given to

183   participants denoted the difference between concrete (i.e., "refers to something that exists

184   in reality") and abstract (i.e., "something you cannot experience directly through your

185   senses or actions") terms. Participants were then asked to rate concreteness of terms using

186   a 1 (*abstract*) to 5 (*concrete*) scale. This data represents a small scale dataset that could be

187   used as pilot data for a study using concrete word ratings. The data is available at

188   https://osf.io/qpmf4/.

189   The second dataset includes a large scale dataset with response latencies, the

₁₉₀ English Lexicon Project (Balota et al., 2007). The English Lexicon Project consists of

₁₉₁ lexical decision response latencies for English words. In a lexical decision task, participants

₁₉₂ simply select "word" for real words (e.g., *dog*) and "nonword" for pseudowords (e.g., *wug*).

₁₉₃ The trial level data is available here: https://elexicon.wustl.edu/. Critically, in each of

₁₉₄ these examples, the individual trial level data for each item is available to simulate and

₁₉₅ calculate standard errors on. Data that has been summarized could potentially be used, as

₁₉₆ long as the original standard deviations for each item were present. From the mean and

₁₉₇ standard deviation for each item, a simulated pilot dataset could be generated for

₁₉₈ estimating new sample sizes. All code to estimate sample sizes is provided on our OSF

₁₉₉ page, and this manuscript was created with a *papaja* (Aust et al., 2022) formatted

₂₀₀ Rmarkdown document.

₂₀₁   For this example, our researcher wants to determine the differences in response

₂₀₂ latencies for abstract and concrete words. They will select $n = 40$ words from the rating

₂₀₃ data that are split evenly into abstract and concrete ends of the rating scale. In the

₂₀₄ experiment, each participant will rate the words for their concreteness, and then complete

₂₀₅ a lexical decision task with these words as the object of interest. Using both datasets, we

₂₀₆ can determine the sample size necessary to ensure adequately measured ratings and

₂₀₇ response latencies.

₂₀₈   *Step 1.* The concreteness ratings data includes 27250 concepts that were rated for

₂₀₉ their concreteness. We randomly selected $n = 20$ abstract words ($M_{Rating} <= 2$) and $n =$

₂₁₀ 20 concrete words ($M_{Rating} >= 4$). In the original study, not every participant rated every

₂₁₁ word, which created uneven sample sizes for each word. Further, participants were allowed

₂₁₂ to indicate they did not know a word, and this data was set to missing data. In our sample

₂₁₃ of 40 words, the average pilot sample size was 28.23 ($SD = 1.35$), and we will use 28 as our

₂₁₄ pilot sample size for this example.

₂₁₅   The ELP response latency data includes 27250 word-forms, 219 that are listed as

₂₁₆ non-words, and 27031 real words. We selected the same words as the concreteness subset

₂₁₇ selected above. The average pilot sample size for this random sample was 32.80 ($SD =$

₂₁₈ 0.56), and $n = 33$ will be our pilot size for this example.

₂₁₉   *Step 2.* Table 2 demonstrates the cutoff scores for deciles of the standard errors for

₂₂₀ the items for the concreteness ratings and lexical decision response latencies. A researcher

₂₂₁ could potentially pick any of these cutoffs or other percentage options not shown here. We

₂₂₂ will use simulation to determine the suggestion that best captures the balance of

₂₂₃ adequately powering our sample and feasibility.

₂₂₄   *Step 3-5.* The pilot data was then simulated, with replacement, with samples from

₂₂₅ 20 to 300 increasing in units of 5, separately for concreteness and lexical decision times

₂₂₆ (Step 3). The standard error of each item was calculated for the bootstrapped samples

₂₂₇ (Step 4), and the percentage of items below each potential cutoff was gathered (Step 5).

₂₂₈ The smallest sample size with at least 80%, 85%, 90%, and 95% of items below the cutoff

₂₂₉ are reported in Table 2 (Step 5).

₂₃₀   *Step 6.* In the last step, the researcher would indicate their smallest sample size, the

₂₃₁ cutoff standard error criterion if they wanted to adaptively test (e.g., examine the SE after

₂₃₂ each participant and stop data collection if all items reached criteria), and their maximum

₂₃₃ sample size. As mentioned earlier, the decile for a balanced standard error cutoff is unclear

₂₃₄ and without guidance, a potential set of researcher degrees of freedom [CITE]. Even

₂₃₅ though both scales appear to converge on similar sample size suggestions for each decile

₂₃₆ and percent level, the impact of scale size (i.e., 1-5 versus 0-2852) and heterogeneity of item

₂₃₇ standard errors (concrete $SD_{SD} = 0.26$ and lexical $SD_{SD} = 116.13$) is not obvious. Last,

₂₃₈ by selecting the ends of the distribution for our concreteness words, skew of the distribution

₂₃₉ may additionally impact our estimates. Each of these will be explored in our simulation.

Simulation Method

In order to evaluate our approach, we used data simulation to create representative pilot datasets of several popular cognitive scales (1-7 measurements, 0-100 percentage measurements, and 0-3000 response latency scale data). For each of these scales, we also manipulated item heterogeneity by simulating small differences in item variances to large differences in item variances based on original scale size. On each of the simulated datasets, we applied the above proposed method to determine how the procedure would perform and evaluate what criteria should be used for cutoff selection (Step 2). This procedure was performed on distributions in the middle of the scale (i.e., normal) and at the ceiling of the scale (i.e., skewed). With this simulation, we will answer several questions:

1) How do pilot data influence sample size suggestions?

    A. How does scale size impact sample size estimations? In theory, the size of the scale used should not impact the power estimates; however, larger scales have a potential for more variability in their item standard deviations (see point C).

    B. How does distribution skew impact sample size estimations? Skew can potentially decrease heterogeneity (i.e., all items are at ceiling, and therefore, variance between item standard errors is low) or could increase heterogeneity (i.e., some items are skewed, while others are not). Therefore, we expect skew to impact the estimates in the same way as point C.

    C. How does heterogeneity impact sample size estimations? Heterogeneity should decrease power (Alexander & DeShon, 1994; Rheinheimer & Penfield, 2001), and thus, increased projected sample sizes should be proposed as heterogeneity of item variances increases.

2) Do the results match what one might expect for traditional power curves? Power

264   curves are asymptotic, that is, they "level off" as sample size increases. Therefore, we

265   expect that our procedure should also demonstrate a leveling off effect as pilot data

266   sample size increases. For example, if one has a 500-person pilot study, our

267   simulations should suggest a point at which items are likely measured well, which

268   may have happened well before 500.

269   3) What should the suggested cutoff standard error decile be?

Data Simulation

271   Table 3 presents the variables and information about the simulations as a summary.

272   *Population.* We simulated data for 30 items using the rnorm function assuming a

273   normal distribution. Each item's population data was simulated with 1000 data points.

274   Items were rounded to the nearest whole number to mimic scales generally collected by

275   researchers. Items were also rounded to their appropriate scale end points (i.e., all items

276   below 0 on a 1-7 scale were replaced with 1, etc.).

277   *Data Scale.* First, the scale of the data was manipulated by creating three sets of

278   scales. The first scale was mimicked after small rating scales (i.e., 1-7 type style) using a μ

279    $= 4$ with a $\sigma = .25$ around the mean to create item mean variability. The second scale

280   included a larger potential distribution of scores with a $\mu = 50$ ($\sigma = 10$) imitating a 0-100

281   scale. Last, the final scale included a $\mu = 1000$ ($\sigma = 150$) simulating a study that may

282   include response latency data in the milliseconds. For the skewed distributions, the item

283   means were set to $\mu = 6$, 85, and 2500 respectively with the same σ values around the item

284   means. Although there are many potential scales, these three represent a large number of

285   potential variables commonly used in the social sciences. As we are suggesting item

286   variances is a key factor for estimating sample sizes, the scale of the data is influential on

287   the amount of potential variance. Smaller data ranges (1-7) cannot necessarily have the

288   same variance as larger ranges (0-100).

289 *Item Heterogeneity.* Next, item heterogeneity was included by manipulating the

290 potential for each individual item. For small scales, the $\sigma = 2$ points with a variability of

291 .2, .4, and .8 for low, medium, and high heterogeneity in the variances between items. For

292 the medium scale of the data, $\sigma = 25$ with a variance of 4, 8, and 16. Finally, for the large

293 scale of the data, $\sigma = 400$ with a variance of 50, 100, and 200 for heterogeneity.

294 *Pilot Data Samples.* Each of the populations shown in Table 3 was then sampled as

295 if a researcher was conducting a pilot study. The sample sizes started at 20 participants

296 per item, increasing in units of 10 up to 100 participants. Each of these samples would

297 correspond to Step 1 of the proposed method where a researcher would use pilot data to

298 start their estimation. Therefore, the simulations included 3 scales X 3 heterogeneity X 2

299 normal/skewed X 9 pilot sample sizes representing a potential Step 1 of our procedure.

300 **Researcher Sample Simulation**

301 In this section, we simulate what a researcher might do if they follow our suggested

302 application of AIPE to sample size planning based on well measured items. Assuming that

303 each pilot sample represents a dataset that a researcher has collected (Step 1), the

304 standard errors for each item were calculated to mimic the AIPE procedure of finding an

305 appropriately small confidence interval, as SE functions as the main component of the

306 formula for normal distribution confidence intervals. Standard errors were calculated at

307 each decile of the items up to 90% (i.e., 0% smallest SE, 10% ..., 90% largest SE). The

308 lower deciles would represent a strict criterion for accurate measurement, as many items

309 would need smaller SEs to meet cutoff scores, while the higher deciles would represent less

310 strict criteria for cutoff scores.

311 We then simulated samples of 20 to 2000 increasing in units of 20 to determine what

312 the new sample size suggestion would be (Step 3). We assume that samples over 500 may

313 be considered too large for many researchers who do not work in teams or have participant

314 funds. However, the sample size simulations were estimated over this amount to determine

315 the pattern of suggested sample sizes (i.e., the function between original sample size and

316 projected sample size).

317   Next, the percentage of items that fall below the cutoff scores and therefore would

318 be considered "well-measured" were calculated for each decile by sample (Step 4). From

319 these data, we pinpoint the smallest suggested sample size at which 80%, 85%, 90% and

320 95% of the items fall below the cutoff criterion (Step 5). These values were chosen as

321 popular measures of "power" in which one could determine the minimum suggested sample

322 size (potentially 80% of the items) and the maximum suggested sample size (selected from

323 a higher percentage, such as 90% or 95%).

324   In order to minimize the potential for random quirks to arise, we simulated the

325 sample selection from the population 100 times and the researcher simulation 100 times for

326 each of those selections. This resulted in 1,620,000 simulations of all combinations of

327 variables (i.e., scale of the data, heterogeneity, data skew, pilot study size, researcher

328 simulation size). The average of these simulations is presented in the results.

Simulation Results

Pilot Data Influence on Sample Size

331   For each variable, the plot of the pilot sample size, projected sample size (i.e., what

332 the simulation suggested), and power levels are presented below. The large number of

333 variables means we cannot plot them all simultaneously, and therefore, we averaged the

334 results across other variables for each plot. The entire datasets can be examined on our

335 OSF page.

Scale Size

337   Figure 1 demonstrates the influence of scale size on the results separate by potential

338 cutoff level. The black dots denote the original sample size for reference. Larger scales have

339 more potential variability, and therefore, we see that percent and millisecond scales project

340 a larger required sample size. This relationship does not appear to be linear with scale size,

341 as percent scales often represent the highest projected sample size. Potentially, this finding

342 is due to the larger proportion of possible variance – the variance of the item standard

343 deviations / total possible variance – was largest for percent scales in this set of simulations

344 ($p_{Percent} = .13$). This finding may be an interaction with heterogeneity, as the Likert scale

345 had the next highest percent variability in item standard errors ($p_{Likert} = .10$), followed by

346 milliseconds ($p_{Milliseconds} = .06$).

Skew

348 Figure 2 displays that ceiling distributions, averaged over all other variables, show

349 slightly higher estimates than normal distributions. This result is consistent across scale

350 type and heterogeneity, as results indicated that they are often the same or slightly higher

351 for ceiling distributions.

Item Heterogeneity

353 Figure 3 displays the results for item heterogeneity for different levels of potential

354 power. In this figure, we found that our suggested procedure does capture the differences in

355 heterogeneity. As heterogeneity increases in item variances, the proposed sample size also

356 increases.

357 Using a regression model, we predicted proposed sample size using pilot sample size,

358 scale size, proportion variability (i.e., heterogeneity), and data type (normal, ceiling). As

359 shown in Table 4, the largest influence on proposed sample size is the original pilot sample

360 size, followed by proportion of variance/heterogeneity, and then data and scale sizes.

Projected Sample Size Sensitivity to Pilot Sample Size

362 In our second question, we examined if the suggested procedure was sensitive to the

363 amount of information present in the pilot data. Larger pilot data is more informative, and

364 therefore, we should expect a lower projected sample size. As shown in each figure

365 presented already, we do not find this effect. These simulations from the pilot data would

366 nearly always suggest a larger sample size - mostly in a linear trend increasing with sample

367 sizes. This result comes from the nature of the procedure - if we base our estimates on a

368 SE cutoff, we will almost always need a bit more people for items to meet those goals. This

369 result does not achieve our second goal.

370   Therefore, we suggest using a correction factor on the simulation procedure to

371 account for the known asymptotic nature of power (i.e., at larger sample sizes power

372 increases level off). For this function in our simulation study, we combined a correction

373 factor for upward biasing of effect sizes (Hedges' correction) with the formula for

374 exponential decay calculations. The decay factor was calculated as follows:

$$1 - \sqrt{\frac{N_{Pilot} - min(N_{Simulation})}{N_{Pilot}}}^{log_2(N_{Pilot})}$$

375   N_{Pilot} indicates the sample size of the pilot data minus the minimum simulated

376 sample size to ensure that the smallest sample sizes do not decay (i.e., the formula zeroes

377 out). This value is raised to the power of log_2 of the sample size of the pilot data, which

378 decreases the impact of the decay to smaller increments for increasing sample sizes. This

379 value is then multiplied by the projected sample size. As shown in Figure 4, this correction

380 factor produces the desired quality of maintaining that small pilot studies should *increase*

381 sample size, and that sample size suggestions level off as pilot study data sample size

382 increases.

Corrections for Individual Researchers

384   We have portrayed that this procedure, with a correction factor, can perform as

385 desired. However, within real scenarios, researchers will only have one pilot sample, not the

386 various simulated samples shown above. What should the researcher do to correct their

387 projected sample size from their own pilot data simulations?

388   To explore if we could recover the new projected sample size from data a researcher

389 would have, we used regression models to create a formula for researcher correction. The

390   researcher employing our procedure would have the possible following variables from their

391   simulations on their (one) pilot dataset: 1) proposed sample size, 2) pilot sample size, 3)

392   estimate of heterogeneity for the items, 4) and the estimate percent of items below the

393   threshold. Given the non-linear nature of the correction, we added each variable and it's

394   non-linear `log2` transform to the regression equation, as this function was used to create

395   the correction. The intercept only model was used as a starting point (i.e., `corrected`

396   `sample ~ 1`), and then all eight variables (each variable and their `log2` transform) were

397   entered into a forward stepwise regression to capture the corrected scores with the most

398   predictive values. Each variable was entered one at a time using the `step` function from

399   the *stats* library in R (R Core Team, 2022).

400   As shown in Table 5, all variables were included in the final equation, each

401   contributing a significant change to the previous model, as defined by ΔAIC > 2 points

402   change between each step of the model. Proposed sample size and original sample size were

403   the largest predictors – unsurprising given the correction formula employed – followed by

404   the percent "power" level and proportion of variance. This formula approximation captures

405    $R^2 = .99$, 90% CI $[0.99, 0.99]$ of the variance in sample size scores and should allow a

406   researcher to estimate based on their own data, $F(8, 4, 527) = 67, 497.54$, $p < .001$. We

407   provide convenience functions in our additional materials to assist researchers in estimating

408   the final adjusted sample size.

Choosing an Appropriate cutoff

410   Last, we examined the question of an appropriate SE decile. First, the 0%, 10%,

411   and 20% deciles are likely too restrictive, providing very large estimates that do not always

412   find a reasonable sample size in proportion to the pilot sample size, scale size, and

413   heterogeneity. If we examine the R^2 values for each decile of our regression equation

414   separately, we find that the values are all $R^2 > .99$ with very little differences between

415   them. Figure 5 illustrates the corrected scores for simulations at the the 40% and 50%

416 decile recommended cutoff for item standard errors. For small heterogeneity, differences in

417 decile are minimal, while larger heterogeneity shows more correction at the 40% decile

418 range, especially for scales with larger potential variance. Therefore, we would suggest the

419 40% decile to overpower each item to determine the minimum and maximum sample size.

420   The final formula for 40% decile correction is provided in Table 6. Proportion of

421 variance can be calculated with the following:

$$\frac{SD_{ItemSD}}{\sqrt{\frac{(Maximum-Minimum)^2}{4}}}$$

422 where maximum and minimum are the max and min values found in the scale (or the data,

423 if the scale is unbounded). This formula would be applied in Step 5 of the proposed

424 procedure. While the estimated coefficients could change given variations on our simulation

425 parameters, the general size and pattern of coefficients was consistent, and therefore, we

426 believe this correction equation should work for a variety of use cases. We will now

427 demonstrate the final procedure on the examples provided earlier.

Updated Example

429   The updated proposal steps are in Table 1. The main change occurs in Step 2 with

430 a designated cutoff decile, and Step 5 with a correction score. Using the data from the 40%

431 decile in Table 2, we can determine that the stopping rule for concreteness ratings would

432 be 0.18, and the stopping rule for lexical decision times would be 51.32. For Step 5, we

433 apply our correction formula separately for each one, as they have different variability

434 scores, and these scores are shown in Table 7. Each row was multiplied by row one's

435 formula, and then these scores are summed for the final sample size. Sample sizes cannot

436 be proportional, so we recommend rounding up to the nearest whole number.

437   For one additional consideration, we calculated the potential amount of data

438 retention given that participants could indicate they did not know a word ($M_{answered} =$

0.90, $SD = 0.13$) in the concreteness task or answer a trial incorrectly in the lexical decision

task ($M_{correct} = 0.78$, $SD = 0.22$). In order to account for this facet, the potential sample

sizes were multiplied by $\frac{1}{p_{retained}}$ where the denominator is proportion retained for each task.

Additional Materials

Package

We have developed functions to implement the suggested procedure as part of an

upcoming package `semanticprimeR`. You can install the package from GitHub using:

`devtools::install_github("SemanticPriming/semanticprimeR")`. We detail the

functions below by proposed step in the process.

Step 1. Ideally, researchers would have pilot data that represented their proposed

data collection. This data should be formatted in long format wherein each row represents

the score from an item by participant, rather than wide format wherein each column

represents an item and each row represents a single participant. The

`tidyr::pivot_longer()` or `reshape::melt()` functions can be used to reformat wide

data. If no pilot data is available, the `simulate_population()` function can be used with

the following arguments (and example numbers, * indicates optional). This function will

return a dataframe with the simulated normal values for each item.

```
# devtools::install_github("SemanticPriming/semanticprimeR")
library(semanticprimeR)
pops <- simulate_population(mu = 4, # item means
  mu_sigma = .2, # variability in item means
  sigma = 2, # item standard deviations
  sigma_sigma = .2, # standard deviation of the standard deviations
  number_items = 30, # number of items
  number_scores = 20, # number of participants
  smallest_sigma = .02, #* smallest possible standard deviation
```

```
    min_score = 1, #* minimum score for truncating purposes

    max_score = 7, #* maximum score for truncating purposes

    digits = 0) #* number of digits for rounding


head(pops)
```

```
## 456    item score

## 457   ## 1    1    3

## 458   ## 2    2    2

## 459   ## 3    3    7

## 460   ## 4    4    2

## 461   ## 5    5    5

## 462   ## 6    6    3
```

Step 2. In step 2, we can use `calculate_cutoff()` to calculate the standard error of the items, the standard deviation of the standard errors and the corresponding proportion of variance possible, and the 40% decile cutoff score. The `pops` dataframe can be used in this function, which has columns named `item` for the item labels (i.e., 1, 2, 3, 4 or characters can be used), and `score` for the dependent variable. This function returns a list of values to be used in subsequent steps.

```
cutoff <- calculate_cutoff(population = pops, # pilot data or simulated data

  grouping_items = "item", # name of the item indicator column

  score = "score", # name of the dependent variable column

  minimum = 1, # minimum possible/found score

  maximum = 7) # maximum possible/found score


cutoff$se_items # all standard errors of items
```

```
##  [1] 0.2926737 0.4376973 0.3966969 0.4639646 0.4308804 0.4729527 0.3973597
```

```
##  [8] 0.3734618 0.3439324 0.3933660 0.3346247 0.3879772 0.4466248 0.3324550
```

```
## [15] 0.4530598 0.4500000 0.4660867 0.3590924 0.2523573 0.4345294 0.4844965
```

```
## [22] 0.4805425 0.4167544 0.3234274 0.2926737 0.3371709 0.3838859 0.4285840
```

```
## [29] 0.3802700 0.2325488
```

```
cutoff$sd_items # standard deviation of the standard errors
```

```
## [1] 0.06798869
```

```
cutoff$cutoff # 40% decile score
```

```
##        40%
```

```
## 0.3824396
```

```
cutoff$prop_var # proportion of possible variance
```

```
## [1] 0.0226629
```

Step 3. The `bootstrap_samples()` function creates bootstrapped samples from the pilot or simulated population data to create samples to estimate the number of participants needed for item standard error to be below the cutoff calculated in Step 2. This function returns a list of samples with sizes that start at the `start` size, increase by `increase`, and end with the `stop` sample size. The population or pilot data will be included in `population`, and the item column indicator should be included in `grouping_items`.

```
samples <- bootstrap_samples(start = 20, # starting sample size
  stop = 100, # stopping sample size
  increase = 5, # increase bootstrapped samples by this amount
  population = pops, # population or pilot data
  replace = TRUE, # bootstrap with replacement?
```

```
  grouping_items = "item") # item column label


head(samples[[1]])
```

```
## # A tibble: 6 x 2
## # Groups:   item [1]
##     item score
##    <int> <dbl>
## 1      1     1
## 2      1     4
## 3      1     1
## 4      1     1
## 5      1     2
## 6      1     5
```

Step 4 and 5. The proportion of bootstrapped items across sample sizes below the cutoff score can then be calculated using `calculate_proportion()`. This function returns a dataframe of each sample size with the proportion of items below that cutoff to use in the next function. The `samples` and `cutoff` were calculated with our previous functions, with the column for item labels and dependent variable to ensure the right calculations.

```
proportion_summary <- calculate_proportion(samples = samples, # samples list
  cutoff = cutoff$cutoff, # cut off score
  grouping_items = "item", # item column name
  score = "score") # dependent variable column name


head(proportion_summary)
```

```
## # A tibble: 6 x 2
```

```
## percent_below sample_size
##           <dbl>        <dbl>
## 1         0.467           20
## 2         0.533           25
## 3         0.967           30
## 4         0.967           35
## 5             1           40
## 6             1           45
```

Step 6. Last, we use the `calculate_correction()` function to correct the sample size scores given the proposed correction formula. The `proportion_summary` from above is used in this function, along with required information about the sample size, proportion of variance from our cutoff calculation, and what power levels should be calculated. Note that the exact percent of items below a cutoff score will be returned, if the values in `power_levels` are not exactly present. The final summary presents the smallest sample size, corrected, for each of the potential power levels.

```
corrected_summary <- calculate_correction(
  proportion_summary = proportion_summary, # prop from above
  pilot_sample_size = 20, # number of participants in the pilot data
  proportion_variability = cutoff$prop_var, # proportion variance from cutoff scores
  power_levels = c(80, 85, 90, 95)) # what levels of power to calculate


corrected_summary
```

```
## # A tibble: 4 x 3
##   percent_below sample_size corrected_sample_size
##           <dbl>        <dbl>                 <dbl>
## 1          96.7           30                  25.3
```

## 2	96.7	30	25.3
## 3	96.7	30	25.3
## 4	96.7	30	25.3

Vignettes

While the examples in this manuscript are traditionally cognitive linguistics focused, any research using repeated items can benefit from newer sampling techniques. Therefore, we provide 12 example vignettes and varied code examples on our OSF page/GitHub site for this manuscript across a range of data types provided by the authors of this manuscript. Examples include psycholinguistics (De Deyne et al., 2008; Heyman et al., 2014; Montefinese et al., 2022), eye tracking data (Ulloa et al., 2014), social psychology (Grahe et al., 2022; Peterson et al., 2022), COVID related data (Montefinese et al., 2021), and cognitive psychology (Barzykowski et al., 2019; Errington et al., 2021; Röer et al., 2013).

Discussion

We proposed and demonstrated examples for a method using AIPE, bootstrapping, and simulation to estimate a minimum and maximum sample size along with a rule for stopping data collection based on narrow confidence intervals on the parameter of interest. We contend that this procedure is specifically useful for studies with multiple items that intend on using item level focused analyses; furthermore, the utility of measuring each item well can extend to many analysis choices. By focusing on collecting quality data, we can suggest that the data is useful, regardless of the outcome of any hypothesis test.

One limitation of these methods would be our decision to use datasets with very large numbers of items to simulate what might happen within one study. For example, the English Lexicon Project includes thousands of items, and by the time we would simulate for all of those, it would likely suggest needing thousands of participants for most items to reach the criterion. Additionally, as the number of items increases, you may also see very small estimates for sample size due to the correction factor (as with large numbers of

items, you could find many items with standard errors below the 40% decile). Therefore, it would be beneficial to consider only simulating what a participant would reasonably complete in a study. Small numbers of repeated items usually result in larger sample sizes proposed from the original pilot data. This result occurs because the smaller number of items means more samples for nearly all to reach the cutoff criteria. These results are not too different from what we might expect for a power analysis using a multilevel model - larger numbers of items tend to decrease necessary sample size, while smaller numbers of items tend to increase sample size.

Second, these methods do not ensure the normal interpretation of power, where you know that you would find a specific effect for a specific test, α, and so on. As discussed in the Introduction, there is not necessarily a one-to-one mapping of hypothesis to analysis, many of the estimations within a traditional power analysis are just that - best approximations for various parameters. These methods could be used together to strengthen our understanding of the sample size necessary for both a hypothesis test and a well-tuned estimation.

Researchers should consider this hybrid approach for AIPE, bootstrapping, and simulation as a tool for hypothesis testing and parameter estimation. The proposed procedure can be beneficial for many different research studies, specifically replication studies, that usually depend on subject sample size but rarely item sample size, in spite of the fact that item sample sizes contribute to language processing (Baayen et al., 2008; Brysbaert & Stevens, 2018). These observed effects can be then replicated and, as a result of several replications, can be applied to meta-analyses. As a result, analysts would be able to use the accuracy and high statistical power to calculate the parameters to assess whether the effect is genuine. This article helps to achieve this goal by encouraging researchers to conduct studies where the power analysis is not based on the size of the effect but on adequate sampling of the stimuli. We argue that this article can be the initial step to apply

AIPE in a manner that can allow researchers to use item information to provide a more accurate and statistically reliable measure of the effect we aimed to investigate. In conclusion, item power analysis is a tool to avoid the waste of resources while ensuring that adequately measured items can be achieved. Well measured data can enable us to counteract the literature that contains false positives, allowing us to achieve replicable, high-quality science to establish answers to scientific questions with precision and accuracy.

References

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin*, *115*(2), 308–314. https://doi.org/10.1037/0033-2909.115.2.308

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science*, *28*(11), 1547–1562. https://doi.org/10.1177/0956797617723724

Aust, F., Barth, M., Diedenhofen, B., Stahl, C., Casillas, J. V., & Siegel, R. (2022). *Papaja: Prepare american psychological association journal articles with r markdown.* https://CRAN.R-project.org/package=papaja

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Barzykowski, K., Niedźwieńska, A., & Mazzoni, G. (2019). How intention to retrieve a memory and expectation that a memory will come to mind influence the retrieval of autobiographical memories. *Consciousness and Cognition*, *72*, 31–48. https://doi.org/10.1016/j.concog.2019.03.011

Batterham, A. M., & Atkinson, G. (2005). How big does my sample need to be? A primer on the murky world of sample size estimation. *Physical Therapy in Sport*, *6*(3), 153–163. https://doi.org/10.1016/j.ptsp.2005.05.004

Beribisky, N. (2019). *A Multi-Faceted Mess: A Review of Statistical Power Analysis in Psychology Journal Articles.*

https://yorkspace.library.yorku.ca/xmlui/handle/10315/36719

Brysbaert, M. (2019). How Many Participants Do We Have to Include in Properly
    Powered Experiments? A Tutorial of Power Analysis with Reference Tables.
    Journal of Cognition, *2*(1), 16. https://doi.org/10.5334/joc.72

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

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for
    40 thousand generally known English word lemmas. *Behavior Research Methods*,
    46(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). LAB: Linguistic
    Annotated Bibliography – a searchable portal for normed database information.
    Behavior Research Methods, *51*(4), 1878–1888.
    https://doi.org/10.3758/s13428-018-1130-8

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable monte carlo
    simulations with the SimDesign package. *The Quantitative Methods for
    Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Chambers, C. D., Feredoes, E., D. Muthukumaraswamy, S., J. Etchells, P., & 1
    Cardiff University Brain Research Imaging Centre, School of Psychology, Cardiff
    University; (2014). Instead of "playing the game" it is time to change the rules:
    Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*,
    1(1), 4–17. https://doi.org/10.3934/Neuroscience.2014.1.4

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A.,
    Ford, C., Volcic, R., & De Rosario, H. (2017). *Pwr: Basic functions for power
    analysis.*

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12),
    1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H.,

   Al-Tamimi, J., Alotaibi, N. E., AlShakhori, M. K., Altmiller, R. M., Arantes, P.,

   Athanasopoulou, A., Baese-Berk, M. M., Bailey, G., Sangma, C. B. A., Beier, E.

   J., Benavides, G. M., Benker, N., BensonMeyer, E. P., . . . Roettger, T. B.

   (2023). Multidimensional signals and analytic flexibility: Estimating degrees of

   freedom in human-speech analyses. *Advances in Methods and Practices in*

   Psychological Science, *6*(3), 25152459231162567.

   https://doi.org/10.1177/25152459231162567

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W.,

   & Storms, G. (2008). Exemplar by feature applicability matrices and other

   Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4),

   1030–1048. https://doi.org/10.3758/brm.40.4.1030

DeBruine, L. (2021). *Faux: Simulation for factorial designs*. Zenodo.

   https://doi.org/10.5281/ZENODO.2669586

Efron, B. (2000). The bootstrap and modern statistics. *Journal of the American*

   Statistical Association, *95*(452), 1293–1296.

   https://doi.org/10.1080/01621459.2000.10474333

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis

   program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11.

   https://doi.org/10.3758/BF03203630

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., &

   Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology.

   eLife, *10*, e71601. https://doi.org/10.7554/eLife.71601

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible

   statistical power analysis program for the social, behavioral, and biomedical

   sciences. *Behavior Research Methods*, *39*(2), 175–191.

   https://doi.org/10.3758/BF03193146

Grahe, J., Chalk, H., Cramblet Alvarez, L., Faas, C., Hermann, A., McFall, J., & Molyneux, K. (2022). EAMMi2 public data. *Open Science Framework.* https://doi.org/10.17605/OSF.IO/X7MP2

Halpern, S. D. (2002). The Continuing Unethical Conduct of Underpowered Clinical Trials. *JAMA*, *288*(3), 358. https://doi.org/10.1001/jama.288.3.358

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. A. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017). *Royal Society Open Science*, *7*(2), 190806. https://doi.org/10.1098/rsos.190806

Heyman, T., De Deyne, S., Hutchison, K. A., & Storms, G. (2014). Using the speeded word fragment completion task to examine semantic priming. *Behavior Research Methods*, *47*(2), 580–606. https://doi.org/10.3758/s13428-014-0496-5

Kelley, K. (2007). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods*, *39*(4), 755–766. https://doi.org/10.3758/BF03192966

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. https://doi.org/10.1037/h0031564

Montefinese, M., Ambrosini, E., & Angrilli, A. (2021). Online search trends and word-related emotional response during COVID-19 lockdown in Italy: a cross-sectional online study. *PeerJ*, *9*, e11858. https://doi.org/10.7717/peerj.11858

Montefinese, M., Vinson, D., Vigliocco, G., & Ambrosini, E. (2022). Italian age of acquisition norms for a large set of words (ItAoA). *Open Science Framework.* https://doi.org/10.17605/OSF.IO/3TRG2

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Nuijten, M. B., Hartgerink, C. H. J., Assen, M. A. L. M. van, Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, *119*(17). https://doi.org/10.1073/pnas.2115228119

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

Rheinheimer, D. C., & Penfield, D. A. (2001). The effects of type i error rate and power of the ANCOVA f test and selected alternatives under nonnormality and variance heterogeneity. *The Journal of Experimental Education*, *69*(4), 373–391. https://doi.org/10.1080/00220970109599493

Röer, J. P., Bell, R., & Buchner, A. (2013). Is the survival-processing memory

advantage due to richness of encoding? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1294–1302. https://doi.org/10.1037/a0031214

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rousselet, G., Pernet, D. C., & Wilcox, R. R. (2022). An introduction to the bootstrap: A versatile method to make inferences by using data-driven simulations. *Meta-Psychology.* https://doi.org/10.31234/osf.io/h8ft7

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., & others. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337356.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Stewart, S., Rinke, E. M., McGarrigle, R., Lynott, D., Lunny, C., Lautarescu, A., Galizzi, M. M., Farran, E. K., & Crook, Z. (2020). *Pre-registration and registered reports: A primer from UKRN.* https://doi.org/10.31219/osf.io/8v2n7

Ulloa, J. L., Marchetti, C., Taffou, M., & George, N. (2014). Only your eyes tell me what you like: Exploring the liking effect induced by other's gaze. *Cognition and Emotion*, *29*(3), 460–470. https://doi.org/10.1080/02699931.2014.919899

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Table 1

Proposed Procedure for Powering Studies with Multiple Items

Step	Proposed Steps	Updated Steps
1	Use representative pilot data.	Use representative pilot data.
2	Calculate standard error of each of the items in the pilot data. Determine the appropriate SE for the stopping rule.	Calculate standard error of each of the items in the pilot data. Using the 40%, determine the cutoff and stopping rule for the standard error of the items.
3	Create bootstrapped samples of your pilot data starting with at least 20 participants up to a maximum number of participants.	Create bootstrapped samples of your pilot data starting with at least 20 participants up to a maximum number of participants.
4	Calculate the standard error of each of the items in the bootstrapped data. From these scores, calculate the percent of items below the cutoff score from Step 2.	Calculate the standard error of each of the items in the bootstrapped data. From these scores, calculate the percent of items below the cutoff score from Step 2.
5	Determine the sample size at which 80%, 85%, 90%, 95% of items are below the cutoff score.	Determine the sample size at which 80%, 85%, 90%, 95% of items are below the cutoff score. Use the correction formula to adjust your proposed sample size based on pilot data size, power, and percent variability.
6	Report all values. Designate one as the minimum sample size, the cutoff score as the stopping rule for adaptive designs, and the maximum sample size.	Report all values. Designate one as the minimum sample size, the cutoff score as the stopping rule for adaptive designs, and the maximum sample size.

Table 2

Sample Size Estimates by Decile for Concreteness Example

Deciles	C Decile SE	C 80	C 85	C 90	C 95	L Decile SE	L 80	L 85	L 90	L 95
Decile 10	0.13	85	85	95	100	39.64	115	135	135	175
Decile 20	0.14	70	75	80	90	45.67	85	90	110	125
Decile 30	0.16	50	55	60	65	46.76	85	90	105	115
Decile 40	0.18	45	50	50	65	51.32	75	75	90	105
Decile 50	0.19	35	35	45	50	71.17	40	45	45	50
Decile 60	0.20	30	35	35	45	78.95	35	35	45	50
Decile 70	0.21	30	30	35	45	87.85	25	35	35	35
Decile 80	0.23	25	30	30	40	100.45	20	25	25	35
Decile 90	0.25	20	20	25	40	121.11	20	20	20	20

Note. Esimates are based on meeting at least the minimum percent of items (e.g., 80%) but may be estimated over that amount (e.g., 82.5%).

Table 3

Parameter Values for Data Simulation

Information	Likert	Percent	Milliseconds
Minimum	1.00	0.00	0.00
Maximum	7.00	100.00	3,000.00
μ	4.00	50.00	1,000.00
Skewed μ	6.00	85.00	2,500.00
σ_μ	0.25	10.00	150.00
σ	2.00	25.00	400.00
Small σ_σ	0.20	4.00	50.00
Medium σ_σ	0.40	8.00	100.00
Large σ_σ	0.80	16.00	200.00

Table 4

Prediction of Proposed Sample Size from Simulated Variables

Term	Estimate	SE	t	p	pr^2
Intercept	-27.30	3.08	-8.87	< .001	.335
Pilot Sample Size	1.51	0.03	54.76	< .001	.951
Scale: Likert v Percent	7.00	1.80	3.89	< .001	.088
Scale: Likert v Milllisecond	25.63	1.87	13.74	< .001	.548
Proportion Variability	312.44	19.86	15.73	< .001	.613
Data: Ceiling v Normal	-7.16	1.41	-5.08	< .001	.142

Table 5

Parameters for All Decile Cutoff Scores

Term	Estimate	SE	t	p	AIC
Intercept	111.049	78.248	1.419	.156	29,996.94
Projected Sample Size	0.429	0.002	185.360	< .001	20,327.79
Pilot Sample Size	-0.718	0.007	-103.787	< .001	14,753.61
Log2 Projected Sample Size	19.522	0.215	90.693	< .001	8,668.73
Log2 Pilot Sample Size	4.655	0.269	17.296	< .001	8,363.69
Log2 Power	-39.367	15.640	-2.517	.012	8,320.82
Proportion Variability	15.434	3.617	4.267	< .001	8,297.71
Log2 Proportion Variability	-0.729	0.232	-3.143	.002	8,289.81
Power	0.606	0.259	2.343	.019	8,286.31

Table 6

Parameters for 40% Decile Cutoff Scores

Term	Estimate	*SE*	*t*	*p*
Intercept	206.589	128.861	1.603	.109
Projected Sample Size	0.368	0.005	71.269	< .001
Pilot Sample Size	-0.770	0.013	-59.393	< .001
Log2 Projected Sample Size	27.541	0.552	49.883	< .001
Log2 Pilot Sample Size	2.583	0.547	4.725	< .001
Log2 Power	-66.151	25.760	-2.568	.010
Proportion Variability	16.405	6.005	2.732	.006
Log2 Proportion Variability	-1.367	0.382	-3.577	< .001
Power	1.088	0.426	2.552	.011

Table 7

*Applied Correction for Each Proposed Sample Size*

| Formula Estimate	Intercept	Projected SS	Pilot SS	Log2 Projected SS	Log2 Pilot SS	Log2 Power	Prop Var	Log2 Prop Var	Power	Corrected Sample	Data...ss
Formula	206.59	0.37	-0.77	27.54	2.58	-66.15	16.40	-1.37	1.09	NA	NA
Concrete 80	1.00	45.00	28.00	5.49	4.81	6.32	0.13	-2.94	80.00	40.25	44.84
Concrete 85	1.00	50.00	28.00	5.64	4.81	6.41	0.13	-2.94	85.00	45.93	51.17
Concrete 90	1.00	50.00	28.00	5.64	4.81	6.49	0.13	-2.94	90.00	45.92	51.15
Concrete 95	1.00	65.00	28.00	6.02	4.81	6.57	0.13	-2.94	95.00	62.14	69.22
LDT 80	1.00	75.00	33.00	6.23	5.04	6.32	0.08	-3.61	80.00	68.47	87.62
LDT 85	1.00	75.00	33.00	6.23	5.04	6.41	0.08	-3.61	85.00	68.13	87.18
LDT 90	1.00	90.00	33.00	6.49	5.04	6.49	0.08	-3.61	90.00	80.88	103.50
LDT 95	1.00	105.00	33.00	6.71	5.04	6.57	0.08	-3.61	95.00	92.80	118.76
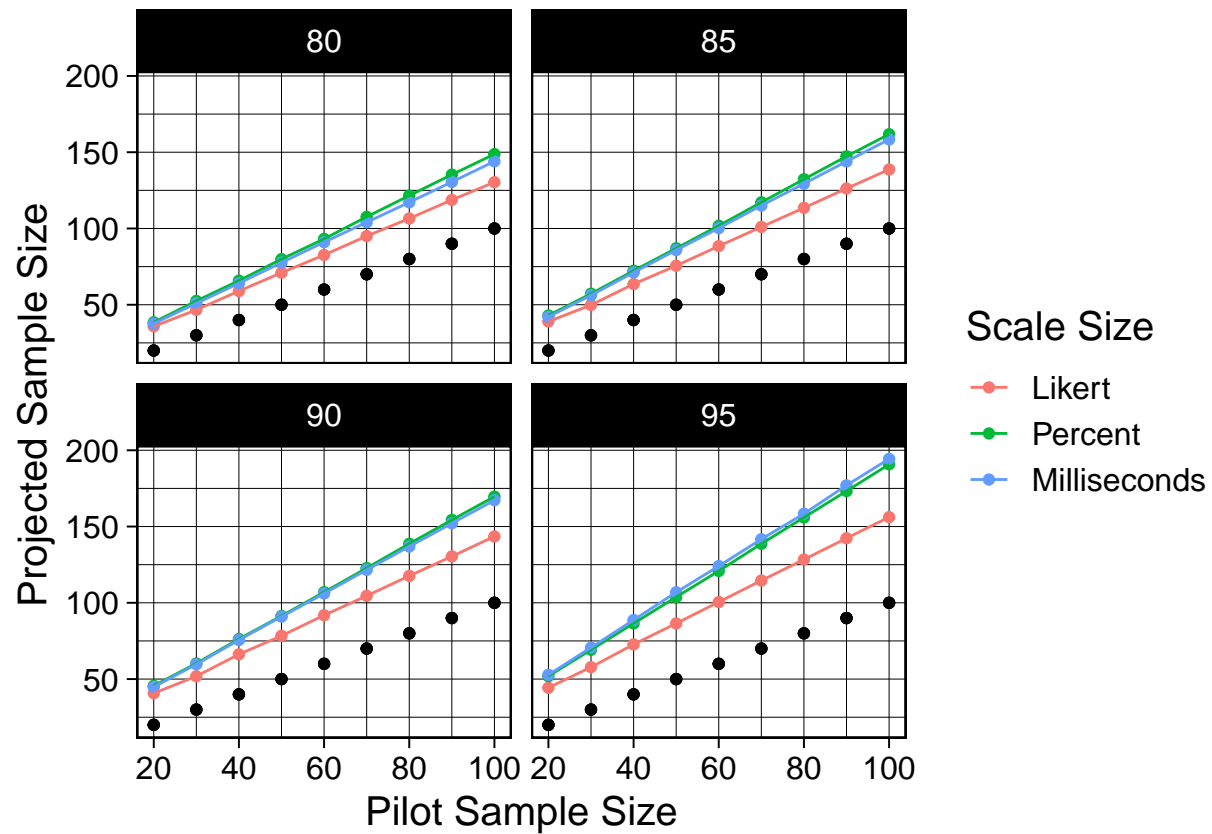
Figure 1

Simulated pilot sample size and final projected sample size to achieve 80%, 85%, 90%, and 95% of items below threshold. These values are averaged over all other variables. Black dots represent original sample size for reference.
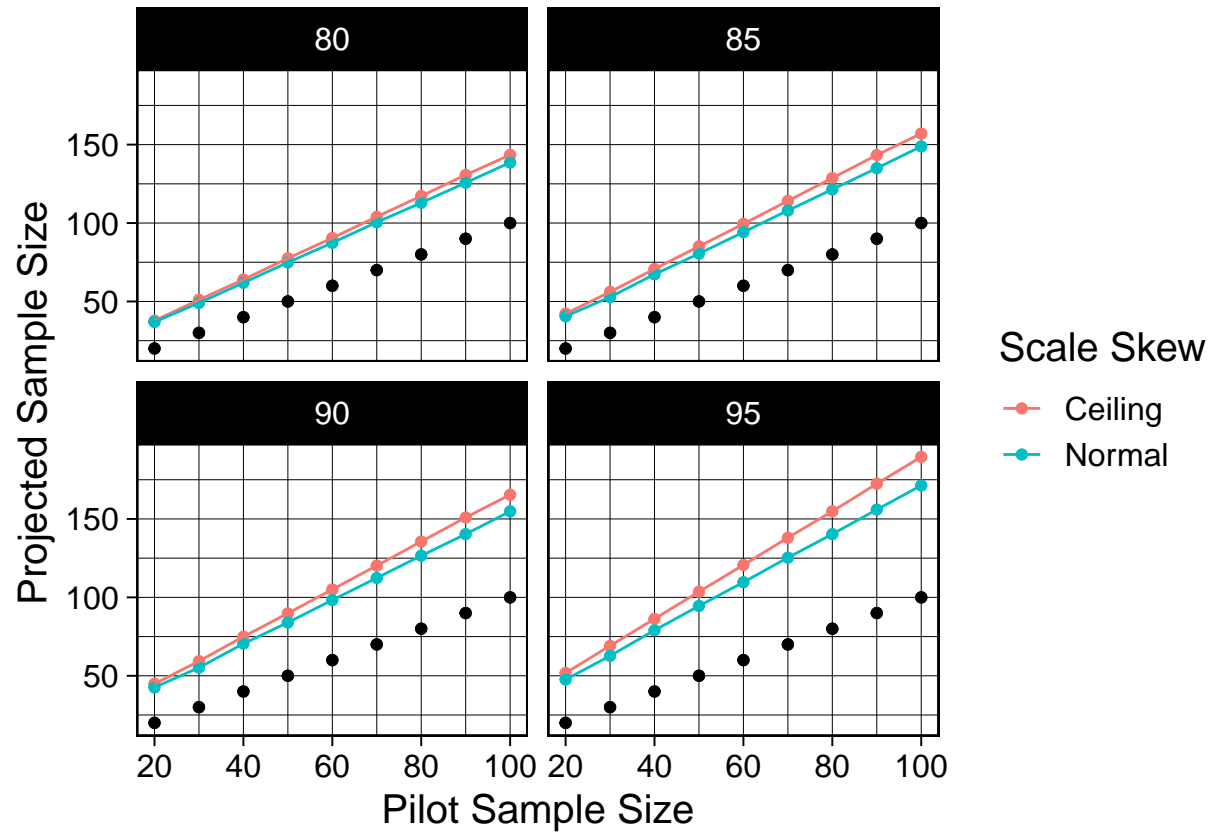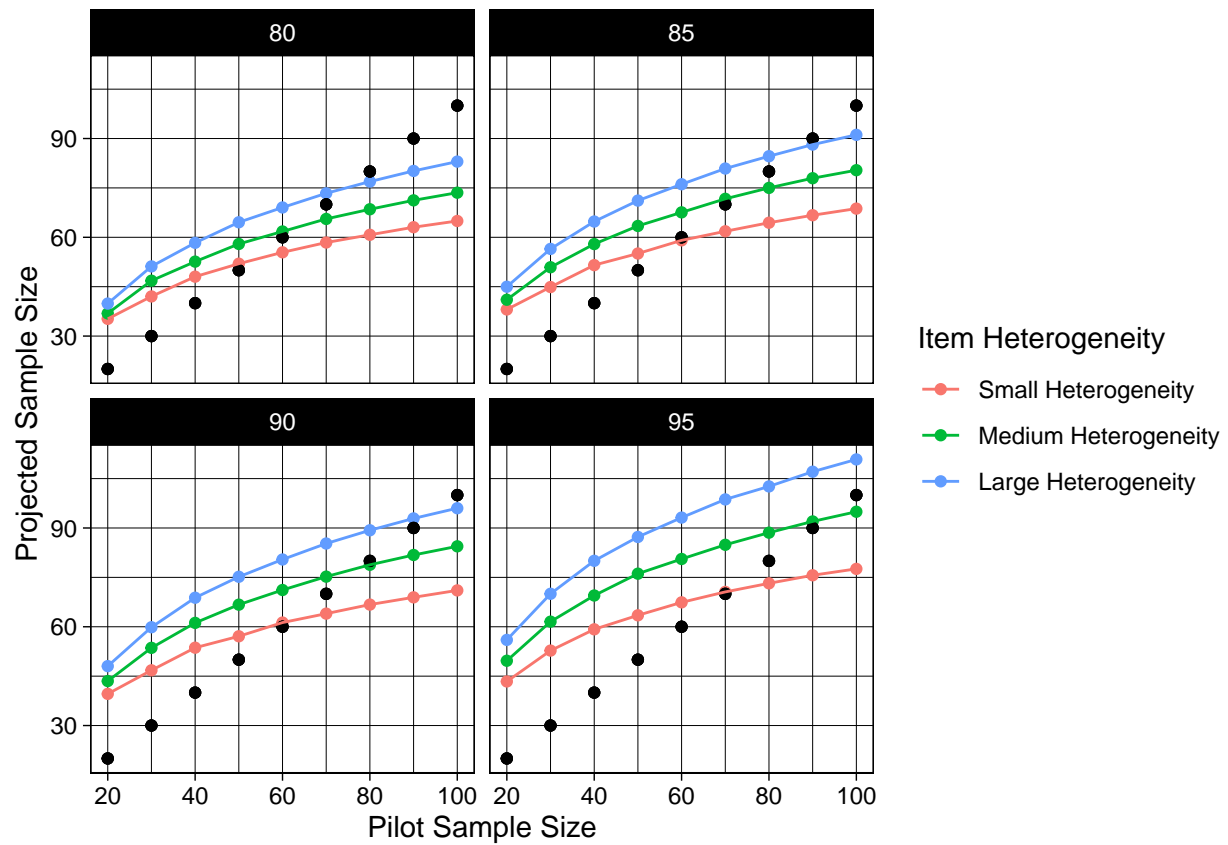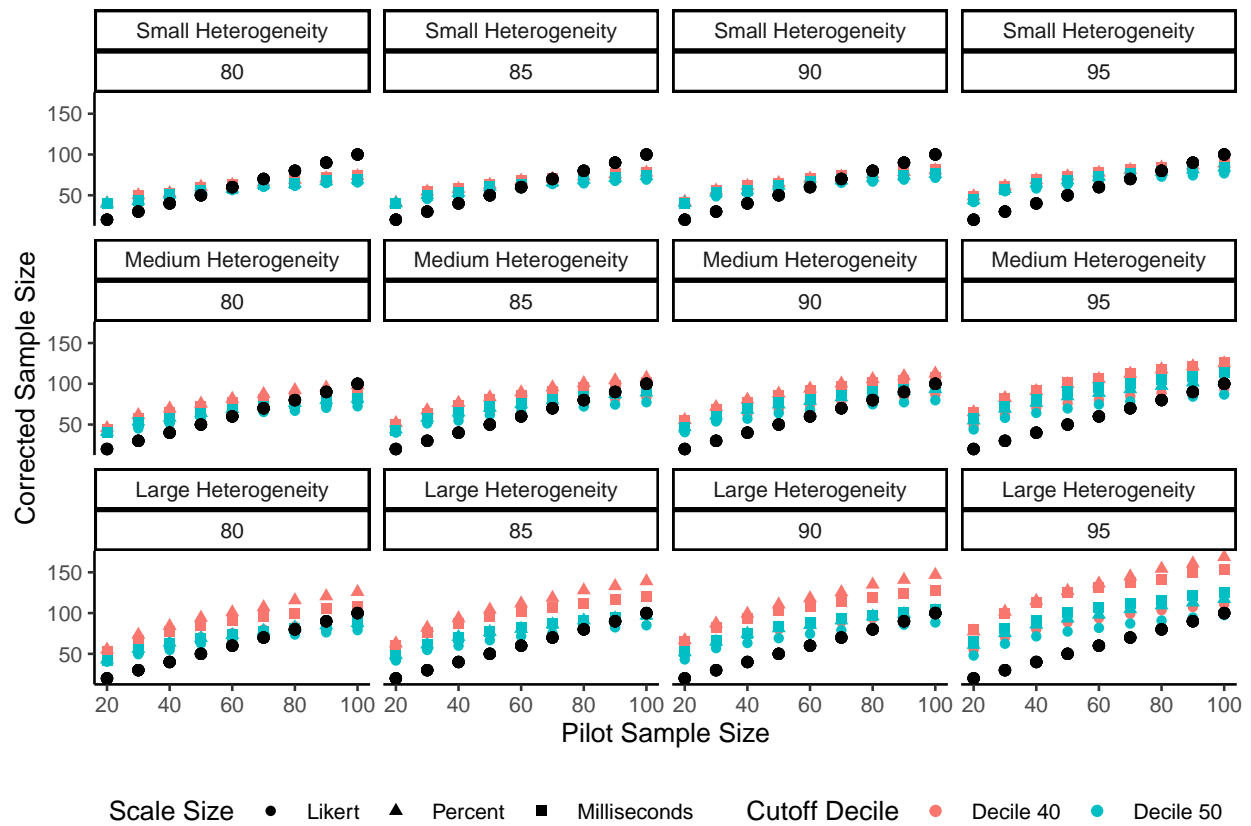
Figure 2

Simulated pilot sample size compared to projected sample size for ceiling versus normal distributions on each scale.
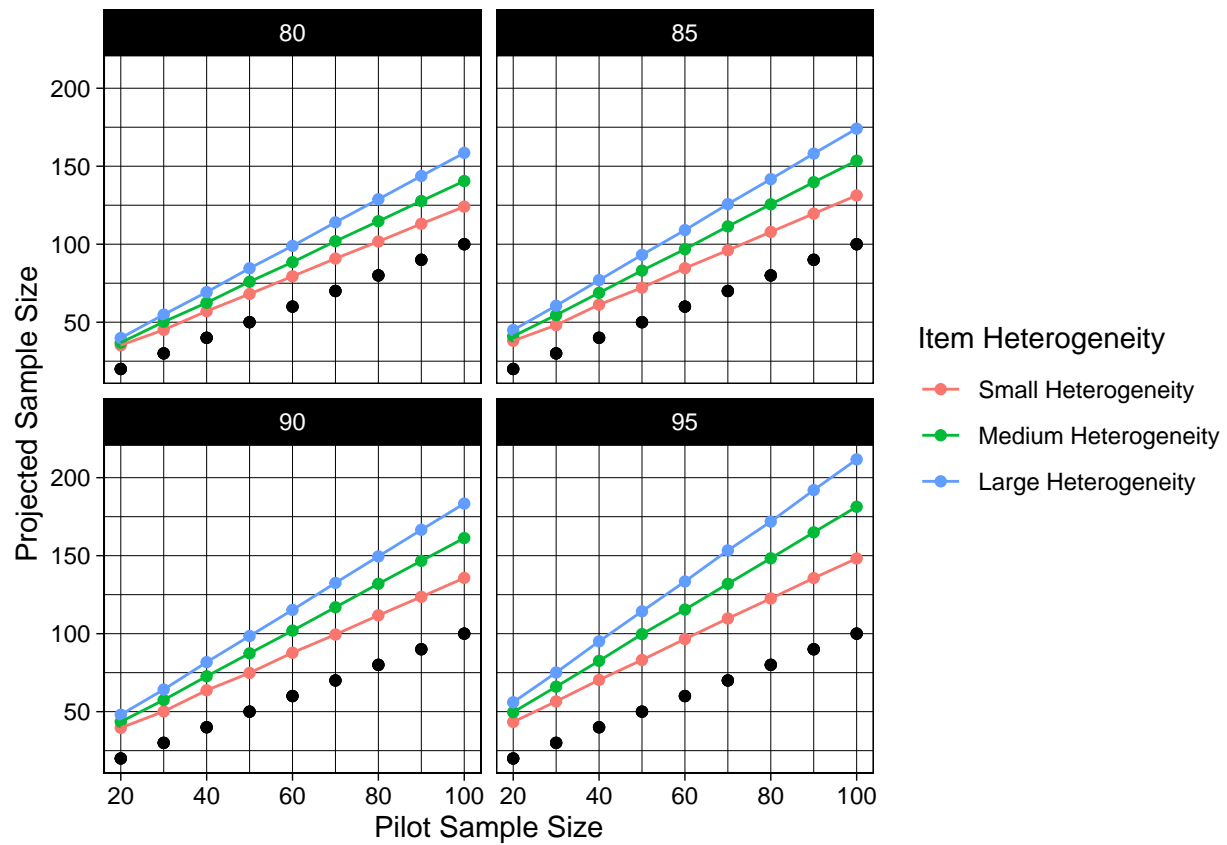
Figure 3

Simulated pilot sample size compared to projected sample size for differing amounts of heterogeneity on each scale.

Figure 4

Corrected projected sample sizes for variability and power levels.

Figure 5

Comparison of the cutoffs for 40% and 50% deciles across heterogeneity (rows), powering of items (columns), and scale size (shape).