Power to the Stimuli: Not the Effect

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

[1] Harrisburg University of Science and Technology

[2] Other Instituions

Author Note

The authors would like to thank K.D. Valentine for her assistance in formulation of correction scores.

The authors made the following contributions. Erin M. Buchanan: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing, Resources, Validation, Visualization, Project Administration, Formal Analysis; Other Folks as Per Order on Doc: Writing - Review & Editing, Data Curation, Resources.

Correspondence concerning this article should be addressed to Erin M. Buchanan, 326 Market St., Harrisburg, PA, 17101. E-mail: ebuchanan@harrisburgu.edu

Abstract

We will add an abstract that explains that we are going to 1) talk about this cool procedure, 2) show some simulations that the procedure works, and 3) give two examples of the procedure in action. You should try it out!

Keywords: power, sampling, accuracy in parameter estimation

Power to the Stimuli: Not the Effect

This part we will put in a real introduction. Things we need to discuss:

- Repeated measures data with items
- Accuracy in Parameter Estimation
- Adaptive sampling
- MLM designs account for item heterogeneity but a lot of people anova

Simulating Sample Size

Using ideas from accuracy in parameter estimation, we suggest the following procedure to determine a sample size for each item:

1) Use pilot data that closely resembles your intended data collection, on the same or similar items that will be used in the study. In this procedure, we will assume that the pilot data is representative of a larger population of sampled items that you intend to assess.
2) Calculate the standard error of each item from the pilot data to create a cutoff score for when an item is "accurately measured". The simulations below will explore what criterion to use when determining the cutoff score from the pilot data.
3) Sample, with replacement, from your pilot data using sample sizes starting at 20 participants and increase in small units (e.g., 20, 25, 30) up to a value that you consider the maximum sample size. We will demonstrate example maximum sample sizes based on the data simulation below; however, a practical maximum sample size may be determined by time (e.g, one semester data collection) or researcher resources (e.g., 200 participants worth of funding). While 20 participants would likely represent an underpowered study, we simply suggest this starting minimum for simulation purposes.
4) For each simulated sample, calculate the standard error for each item, and use these

values to ascertain the percentage of items that meet the cutoff score determined in step 2.

5) Find the minimum sample size that meets 80%, 85%, 90%, and 95% of the items. We recommend these scores to ensure that most items are accurately measured, in a similar vein to common power criteria suggestions. Each researcher can determine which of these is their minimum or maximum sample size (e.g., individual can choose to use 80% as a minimum and 90% as a maximum).
6) Report these values, and designate a minimum sample size, the cutoff criterion, and the maximum sample size. Each researcher should also report if they plan to use an adaptive design, which would stop data collection after meeting the cutoff criterion for each item.

Key Issues

Given the long history of research on power, there are a few key issues that this procedure should address:

1) We should see differences in projected sample sizes based on the variability in the variance for those items (i.e., heterogeneity should increase projected sample size).
2) We should see projected sample sizes that "level off" when pilot data increases. As with regular power estimates, studies can be "overpowered" to detect an effect, and this same idea should be present. For example, if one has a 500 person pilot study, our simulations should suggest a point at which items are likely measured well, which may have happened well before 500.

Method

Data Simulation

Population. The data was simulated using the `rnorm` function assuming a normal distribution for 30 scale type items. Each population was simulated with 1000 data points.

No items were rounded for this simulation.

First, the scale of the data was manipulated by creating three sets of scales. The first scale was mimicked after small rating scales (i.e., 1-7 type style) using a $\mu = 4$ with a $\sigma = .25$ around the mean to create item mean variability. The second scale included a larger potential distribution of scores with a $\mu = 50$ ($\sigma = 10$) imitating a 0-100 scale. Last, the final scale included a $\mu = 1000$ ($\sigma = 150$) simulating a study that may include response latency data in the milliseconds. While there are many potential scales, these three represent a large number of potential variables in the social sciences. As we are suggesting item variances as a key factor for estimating sample sizes, the scale of the data is influential on the amount of *potential* variance. Smaller ranges of data (1-7) cannot necessarily have the same variance as larger ranges (0-100).

Next, item variance heterogeneity was included by manipulating the potential σ for each individual item. For small scales, the $\sigma = 2$ points with a variability of .2, .4, and .8 for low, medium, and high heterogeneity in the variances between items. For the medium scale of data, $\sigma = 25$ with a variance of 4, 8, and 16. Last, for the large scale of data, $\sigma = 400$ with a variance of 50, 100, and 200 for heterogeneity.

Samples. Each population was then sampled as if a researcher was conducting a pilot study. The sample sizes started at 20 participants per item increasing in units of 5 up to 100 participants.

Cutoff Score Criterions. The standard errors of each item were calculated to mimic the AIPE procedure of finding an appropriately small confidence interval, as standard error functions as the main component in the formula for normal distribution confidence intervals. Standard errors were calculated at each decile of the items up to 90% (i.e., 0% smallest SE, 10% ..., 90% largest SE). The lower deciles would represent a strict criterion for accurate measurement, as many items would need smaller SEs to meet cutoff scores,

while the higher deciles would represent less strict criterions for cutoff scores.

Researcher Sample Simulation

In this section, we simulate what a researcher might do if they follow our suggested application of AIPE to sample size planning based on well measured items. Assuming each pilot sample represents a dataset a researcher has collected, we will simulate samples of 20 to 500 to determine what the new sample size suggestion would be. We assume that samples over 500 may be considered too large for many researchers who do not work in teams or have participant funds. The standard error of each item was calculated for each suggested sample size by pilot sample size by population type.

Next, the percent of items that fall below the cutoff scores, and thus, would be considered "well-measured" were calculated for each decile by sample. From this data, we pinpoint the smallest suggested sample size at which 80%, 85%, 90%, and 95% of the items fall below the cutoff criterion. These values were chosen as popular measures of "power" in which one could determine the minimum suggested sample size (potentially 80% of items) and the maximum suggested sample size (potentially 90%).

Results

Differences in Item Variance

We examined if this procedure is sensitive to differences in item heterogeneity, as we should expect to collect larger samples if we wish to have a large number of items reach a threshold of acceptable variance; potentially, assuring we *could* average them if a researcher did not wish to use a more complex analysis such as multilevel modeling.

Figure 1 illustrates the potential minimum sample size for 80% of items to achieve a desired cutoff score. The black dots denote the original sample size against the suggested sample size. By comparing the facets, we can determine that our suggested procedure does

capture the differences in heterogeneity. As heterogeneity increases in item variances, the proposed sample size also increases, especially at stricter cutoffs. Missing cutoff points where sample sizes proposed would be higher than 500.

Projected Sample Size Sensitivity to Pilot Sample Size

In our second question, we examined if the suggested procedure was sensitive to the amount of information present in the pilot data. Larger pilot data is more informative, and therefore, we should expect a lower projected sample size. As shown in Figure 2 for only the low variability and small scale data, we do not find this effect. These simulations from the pilot data would nearly always suggest a larger sample size - mostly in a linear trend increasing with sample sizes. This result comes from the nature of the procedure - if we base our estimates on a SE cutoff, we will almost always need a bit more people for items to meet those goals. This result does not achieve our second goal.

Therefore, we suggest using a correction factor on the simulation procedure to account for the known asymptotic nature of power (i.e., at larger sample sizes power increases level off). For this function in our simulation study, we combined a correction factor for upward biasing of effect sizes (Hedges' correction) with the formula for exponential decay calculations. The decay factor was calculated as follows:

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

N_{Pilot} indicates the sample size of the pilot data minus the minimum simulated sample size to ensure that the smallest sample sizes do not decay (i.e., the formula zeroes out). This value is raised to the power of log_2 of the sample size of the pilot data, which decreases the impact of the decay to smaller increments for increasing sample sizes. This value is then multiplied by the projected sample size. As show in Figure 3, this correction factor produces the desired quality of maintaining that small pilot studies should *increase*

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

Corrections for Individual Researchers

We have portrayed that this procedure, with a correction factor, can perform as desired. However, within real scenarios, researchers will only have one pilot sample, not the various simulated samples shown above. What should the researcher do to correct their projected sample size from their own pilot data simulations?

To explore if we could recover the new projected sample size from data a researcher would have, we used linear models to create a formula for researcher correction. First, the corrected projected sample size was predicted by the original projected sample size. Next, the standard deviation of the item standard deviations was added to the equation to recreate heterogeneity estimates. The scale of the data is embedded into the standard deviation of the items ($r = 0.81$), and therefore, this variable was not included separately. Last, we included the pilot sample size.

The first model using pilot sample size to predict new sample size was significant, $F(1, 5666) = 48{,}042.64$, $p < .001$, $R^2 = .89$, 90% CI $[0.89, 0.90]$, capturing nearly 90% of the variance, $b = 0.62$, 95% CI $[0.62, 0.63]$. The second model with item standard deviation was better then the first model $F(1, 5665) = 55.21$, $p < .001$, $R^2 = .89$, 90% CI $[0.89, 0.90]$. The item standard deviation predictor was significant, $b = 0.02$, 95% CI $[0.01, 0.03]$, $t(5665) = 4.54$, $p < .001$. The addition of the original pilot sample size was also significant, $F(1, 5664) = 9{,}529.83$, $p < .001$, $R^2 = .96$, 90% CI $[0.96, 0.96]$.

As shown in the final model Table 1, the new suggested sample size is proportional to the original suggested sample size (i.e., $b < 1$), which reduces the sample size suggestion. As variability increases, the suggested sample size also increases to capture differences in heterogeneity shown above; however, this predictor is not significant in the final model, and

only contributes a small portion of overall variance. Last, in order to correct for large pilot data, the original pilot sample size decreases the new suggested sample size. This formula approximation captures 96% of the variance in sample size scores and should allow a researcher to estimate based on their own data.

Choosing an Appropriate Cutoff

Last, we examine the question of an appropriate SE decile. All graphs for power, heterogeneity, scale, and correction are presented online. First, the 0%, 10%, and 20% deciles are likely too restrictive, providing very large estimates that do not always find a reasonable sample size in proportion to the pilot sample size, scale, and heterogeneity. If we examine the R^2 values for each decile of our regression equation separately, we find that the 50% (0.96) represents the best match to our corrected sample size suggestions. The 50% decile, in the corrected format, appears to meet all goals: 1) increases with heterogeneity and scale of data, and 2) higher suggested values for small original samples and a leveling effect at larger pilot data. Figure 4 illustrates the corrected scores for simulations at the 50% decile recommended cutoff for item standard errors.

The formula for finding the corrected sample size using a 50% decile is: $N_{CorrectedProjected} = 39.269 + 0.700 \times X_{N_{Projected}} + 0.003 \times X_{SDItems} - 0.694 \times X_{N_{Pilot}}$. The suggested sample size will be estimated from the 80%, 85%, 90%, or 95% selection at the 50% decile as shown above. The item SD can be calculated directly from the data, and the pilot sample size is the sample size of the data from which a researcher is simulating their samples. Therefore, we will recommend the 50% decile of the item standard errors for step 2 of our suggested simulation procedure, and to correct the projected sample sizes found in step 5 using the correction equation above. While the estimated coefficients could change given variations on our simulation parameters, the general size and pattern of coefficients was consistent, and therefore, we believe this correction equation should work for a variety of use cases.

Examples

In this section, we provide two examples of the suggested procedure. The first example includes concreteness ratings from Brysbaert et al. [CITE, 2014]. Instructions given to participants denoted the difference between concrete (i.e., "refers to something that exists in reality") and abstract (i.e., "something you cannot experience directly through your senses or actions") terms. Participants were then asked to rate concreteness of terms using a 1 (abstract) to 5 (concrete) scale. This data represents a small scale dataset that could be used as pilot data for a study using concrete word ratings. The data is avaliable at OSF LINK [CITE]. The second dataset includes a large scale dataset with response latencies, the English Lexicon Project [CITE, Balota et al.]. The English Lexicon Project consists of lexical decision response latencies for English words. In a lexical decision task, participants simply select "word" for real words (e.g., *dog*) and "nonword" for pseudowords (e.g., *wug*). The trial level data is avaliable here [https://elexicon.wustl.edu/, CITE]. Critically, in each of these examples, the individual trial level data for each item is available to simulate and calculate standard errors on. Data that has been summarized could potentially be used, as long as the original standard deviations for each item were present. From the mean and standard deviation for each item, a simulated pilot dataset could be generated for estimating new sample sizes. All code to estimate sample sizes is provided on our OSF page.

Concreteness Ratings

The concreteness ratings data includes 63039 concepts that were rated for their concreteness. In our fictional study for this example, we selected 100 random words to show participants. In the original study, not every particpant rated every word, which created uneven sample sizes for each word. In our random sample of 100 words, the average pilot sample size was 27.96 ($SD = 1.47$), and we will use 28 as our pilot sample size for this example. All "do not know" ratings were set as missing data. The 50% decile for

items standard error was 0.25 for our cutoff criterion.

The pilot data was then simulated, with replacement, with samples from 20 to 300 increasing in units of 5. On each sample, the percent of items below the cutoff score were calculated. After applying our correction equation, we find that a sample size of 44 would allow for at least 80% of items to meet the cutoff criterion. The sample sizes for 85% (48), 90% (48), and 95% (51) are also options for sample size suggestions. Finally, we calculated the potential amount of data retention given that participants could indicate they did not know a word ($M_{correct} = 0.82$, $SD = 0.24$). In order to account for this facet, the potential sample sizes were multiplied by 1/0.82, which results in a suggested sample of 54, 59, and 63. Therefore, we could designate our minimum sample per item as 54, stopping rule of 0.25, and maximum sample size of 63.

Response Latencies

The ELP response latency data includes 80962 word-forms, 40481 that are listed as non-words, and 40481 real words. For our example study, we will randomly select 500 real words and 500 non-words to show participants. The average pilot sample size for this random sample was 32.68 ($SD = 0.64$), and $n = 33$ will be our pilot size for this example. Again, participants are expected to make mistakes, and we calculated percent correct as 0.86, which was roughly even in the two stimulus categories: $M_{word} = 0.84$ and $M_{non-word} = 0.87$. The 50% decile for items standard error was 61.32 for our cutoff criterion. We additionally checked to ensure that the two stimulus types did not have very different cutoff criterions: 50% decile $SE_{words} = 56.94$, 50% decile $SE_{nonwords} = 65.61$. In this scenario, we could chose to go with the lower SE to be more conservative (i.e., higher projected sample size). Given the values were close for large scale data, we used the 50% decile of all stimuli taken together.

The pilot response latency data was then simulated in the same way as described

above. After calculating the percent below our cutoff score, we applied the correction to the projected sample sizes. A sample size of 31 would equate to 80% of the items reaching our cutoff, along with 85% (34), 90% (34), and 95% (38). Again, we adjusted for data loss given that participants are expected to incorrectly answer items, resulting in a suggested sample of 36, 40, and 44. One other possible consideration for this study is potential fatigue in showing participants 1000 target items. Therefore, we could designate in our research design that each participant will only receive 500 of the target items. We would need to double our sample sizes to account for splitting of the items across multiple sets of participants. Our minimum sample size for the entire study could be 72, stopping rule of 61.32, and maximum sample size of 88. This study would benefit from an adaptive design, where smaller sets items are randomly sampled for participants until they reach the minimum sample size or the cutoff criteria. At this point, items are probablistically sampled (e.g., higher selection probability for items that have not reached a minimum or stopping rule) until all items have reached criteria.

Additional Materials

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 XX example vignettes on our OSF page/GitHub site for this manuscript across a range of examples of data types provided by the authors of this manuscript. Examples include ... ADD HERE AFTER DONE WITH VIGNETTES.

Discussion

To be added once people say this idea doesn't suck.

References

Table 1

Parameters for All Decile Cutoff Scores

Term	Estimate	*SD*	*t*	*p*
Intercept	39.269	0.437	89.843	$< .001$
Projected Sample Size	0.700	0.002	366.669	$< .001$
Item SD	0.003	0.003	0.952	.341
Pilot Sample Size	-0.695	0.007	-97.621	$< .001$

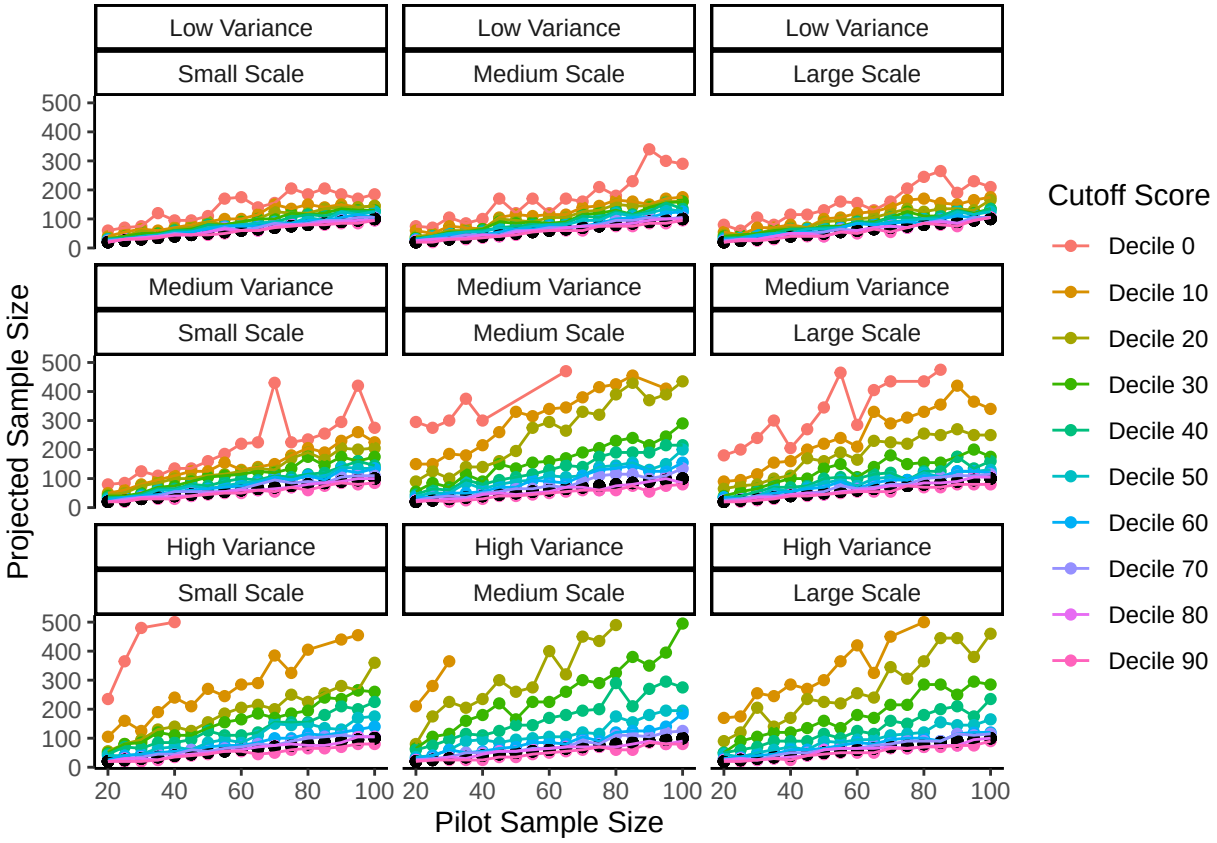

Figure 1. Add a good caption here.

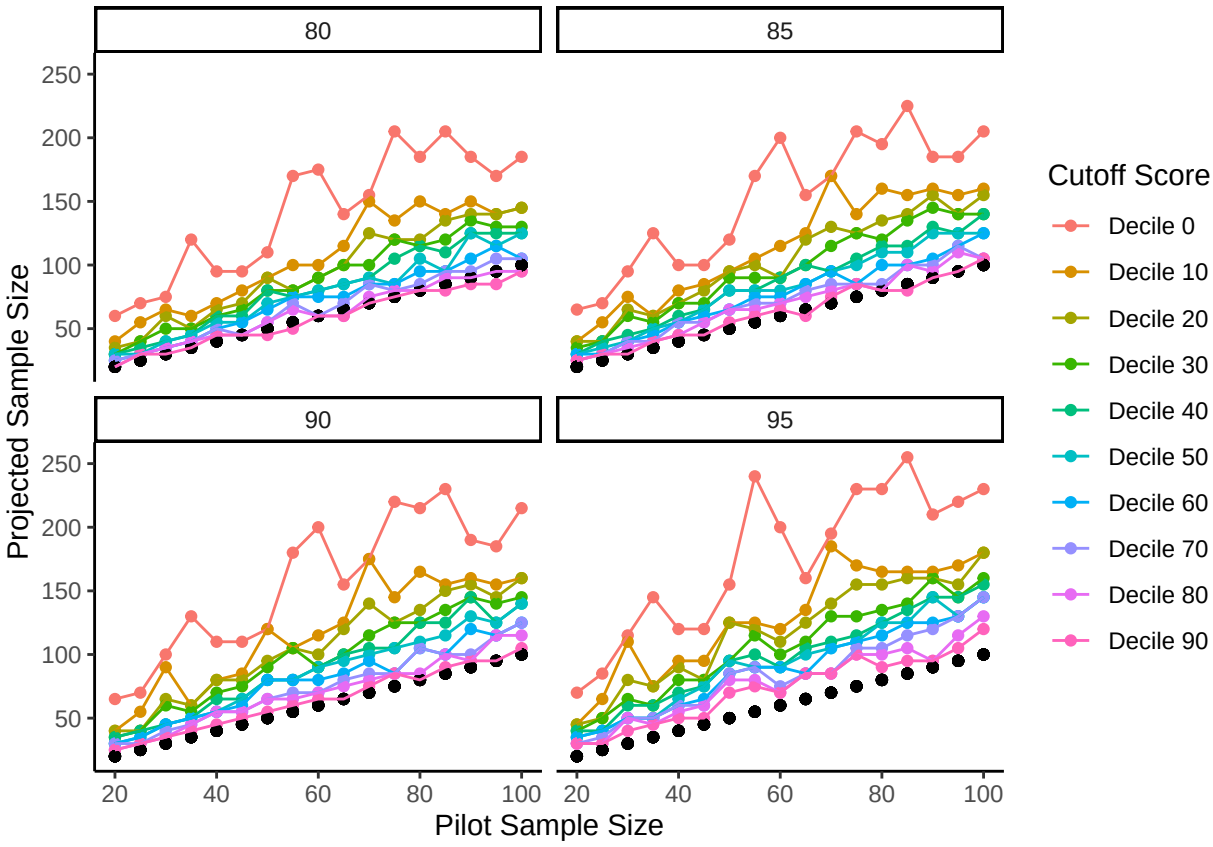

Figure 2. Add good description here.

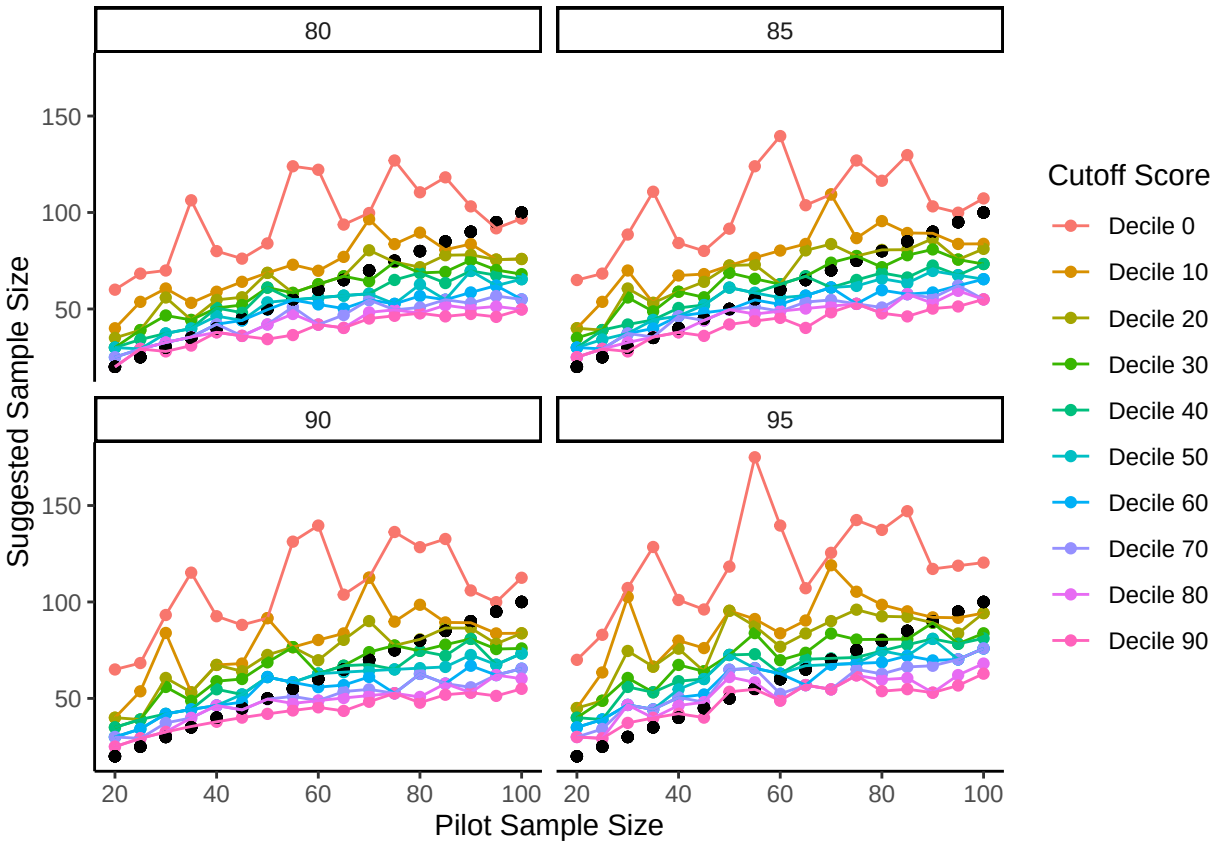

Figure 3. A corrected figure update this caption.

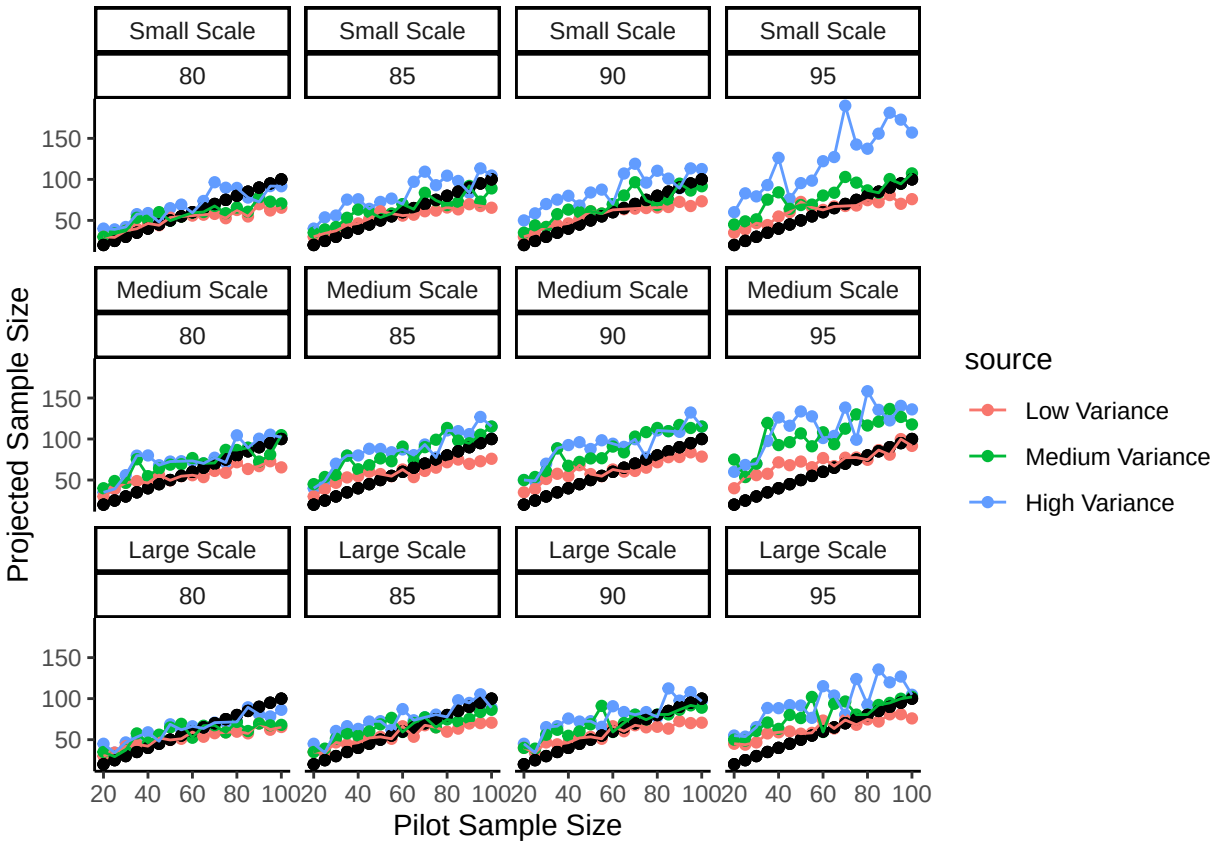

Figure 4. A picture of the 50% cutoff.