Stage 1 Registered Report: Can we shift belief in the ‘Law of Small Numbers?’

## Abstract

Even statistically-trained people show ‘sample size neglect,’ i.e. they underestimate how much the variability of mean estimates increases as sample size decreases. We consider whether judging whether simulated datasets come from different populations can help people overcome this bias. Before and after training with simulated datasets, participants are given parallel forms of a 12-item quiz testing statistical knowledge, including six items testing sample size neglect. We hypothesise that the training will counteract sample size neglect in some participants, and those who show learning will also show improvement on quiz items relating to sample size neglect.

#### Keywords: statistical reasoning, power, online training, sample size neglect

## Introduction

Compared to laypeople, scientists receive extensive training to help them understand and appropriately address the uncertainty of evidence. Yet many scientists fall short in their understanding of statistical concepts. One cognitive bias demonstrated by Tversky and Kahneman (1971) is the ‘belief in the law of small numbers.’ This refers to the tendency to overestimate the stability of estimates that come from small samples - which, following Yoon, Scopelliti, and Morewedge (2021), we shall term ‘sample size neglect.’ For instance, as shown in Figure 1, if repeatedly sampling 10 men from a population, the mean height of the sample will be far more variable than when repeatedly sampling 60 men. People understand that sample size does not affect the expected mean value, but they tend not to appreciate that it has a large effect on the standard error of the mean (i.e. variability of the red bars). This has implications for understanding of statistical power, i.e. the numerical relationship between sample size and ability to detect a true effect. Sample size neglect can help explain why so many studies in psychology, and indeed many other scientific disciplines, are underpowered.

# Method

## Participants

Criteria for participants are: (1) age of 18 years or over, (2) have studied life or social sciences for at least one term at undergraduate (bachelor’s degree) level. We will recruit up to 100 participants via the online research platform Prolific (www.prolific.co) and social media platforms. The maximum sample size is determined from a power calculation (see section on Simulated data for sample size determination, below) indicating that this is sufficient to detect the case when 33% or more of participants improve their performance on the training task.

### Independent and dependent variables

On the training task, the independent variable is block: there are 4 blocks each of 20 items, but our focus is on comparing blocks 1 and 4. The dependent variables are (a) earnings in the game, which reflects both proportion correct for each block, and ability to select the optimal array index at which to respond (see below); (b) mean array index per block, 1 to 6, corresponding to the array size (10, 20, 40, 80, 160, 320 per group) at which the response is made. The earnings score is highly correlated with percentage correct (r = .989).

## References

Verhoef, E., Shapland, C. Y., Fisher, S. E., Dale, P. S., & Pourcain, B. S. (2021). The developmental origins of genetic factors influencing language and literacy: Associations with early-childhood vocabulary. *Journal of Child Psychology and Psychiatry*, *62*(6), 728–738. <https://doi.org/10.1111/jcpp.13327>