This postprint was originally published by Cambridge University Press as:
Gigerenzer, G. (2022). We need to think more about how we conduct research. *Behavioral and Brain Sciences, 45*, Article e16. https://doi.org/10.1017/S0140525X21000327.

We need to think more about how we conduct research

Gerd Gigerenzer

Abstract

Research practice is too often shaped by routines rather than reflection. The routine of sampling subjects, but not stimuli, is a case in point, leading to unwarranted generalizations. It likely originated out of administrative rather than scientific concerns. The routine of sampling subjects and testing their averages for significance is reinforced by delusions about its meaningfulness, including the replicability delusion.

The replicability crisis has made us rethink research practice. Should we lower the level of significance from 0.05 to 0.005, replace *p*-values with Bayes-factors, or require preregistration? Yarkoni rightly advocates looking even deeper into what fuels unwarranted generalizations.

Select or sample stimuli?

As Yarkoni notes, the routine practice is to sample subjects but not stimuli, although the choice of stimuli can more strongly influence a result. This happens if individuals are more alike than stimuli or if the stimuli are not representative – akin to a survey reporting only the extreme opinions of a few selected people rather than those of a representative sample (Brunswik, 1956). Consider two prominent cases where generalizations based on selected stimuli become invalid.

Take the claim that people are overconfident. It is based, among others, on a large number of studies asking general knowledge questions such as "Which city lies further south: Rome or New York? How confident are you?" On average, confidence was higher than proportion correct. Rome, however, is further north than New York yet warmer. Selecting unusual stimuli generates a semblance of general overconfidence. When stimuli were instead randomly sampled from a population (such as all large cities in the world), average confidence matched proportion correct (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, Winman, & Olssen, 2000).

The same holds for the generalization that availability makes people overestimate the likelihood that a letter more likely occurs first than third in a word. In the original study (Tversky & Kahneman, 1973), two-thirds overestimated this for the five consonants selected – all of which more likely appear in the third position, which is untypical. But when representative samples of all letters were used, judgments corresponded to the actual letter frequencies rather than to availability (Sedlmeier, Hertwig, & Gigerenzer, 1998).

In both cases, the sampling of stimuli, not subjects, made the essential difference. By picking unusual stimuli, one can produce results that do not generalize.

Why sample subjects but not stimuli?

One might object that sampling subjects is, and has always been, the method of experimental psychology since its beginnings in Wundt's laboratory. In fact, research practice consisted of careful analysis of single individuals exposed to many stimuli. Wundt himself served as a subject, tested by a technician on a range of stimuli. Luria studied the mind of the mnemonist Shereshevsky using a broad range of stimuli, including words, numbers, and tones. Skinner studied one pigeon at a time, reporting cumulative records instead of averages. Simon studied individual chess players, varying chess positions.

One might also object that sampling subjects is required by inferential statistics. That is also incorrect. Take Fisher's *Design of Experiments* (1935), which introduced psychologists to randomized experiments, null hypotheses, and significance. It reported one psychological experiment involving a single subject (a lady who allegedly knew whether tea or milk was first poured into a cup) and a sample of stimuli (cups of tea). Nowhere in Fisher's experiments were subjects ever sampled (Gigerenzer, 2006). Why did research practice change?

In his seminal book *Constructing the Subject*, Danziger (1990) argues that American psychologists' reason for abandoning carefully study of individuals in the 1930s and 1940s and embracing averages as their new "subject" had little to do with science. They reacted to university administrators' pressure to prove that their research was useful for applied fields, specifically educational research, which offered large sources of funding. The use of averages in treatment groups (such as pupils) was quickly adopted in educational psychology and parapsychology, whilst the core of scientific psychology, such as research on perception research, continued to conduct studies with few individuals. If Danziger is right, our current research practice – sampling subjects only, and testing their averages for significance – is a historical artifact motivated by the needs of administrators, not science.

The replication delusion

Routines easily turn into blind spots, which also exist in the sampling of subjects. The statistical theory underlying significance testing assumes that random samples are drawn from a population, yet most researchers do not sample subjects (or stimuli) randomly from a population or define a population in the first place. Without meeting the assumptions of the model, one cannot know the population to which a significant result might generalize, or where it might be replicated. That fundamental mismatch between statistical theory and experimental practice is bridged by a delusion:

> A significant result $p = 0.01$ logically implies that if the experiment were repeated a large number of times, one would expect to obtain a significant result on 99% of occasions.

This delusion provides unwarranted certainty that a result is replicable, and makes replication studies appear obsolete – and not worth publishing. How widespread is it? In all existing studies, mostly conducted in the last decade, 839 academic psychologists in Chile, Germany, Italy, the Netherlands, Spain, and the UK had been asked to evaluate the above statement – the replication delusion – including variants such as $p = 0.001$, which made no difference (Gigerenzer, 2018) (Fig. 1).

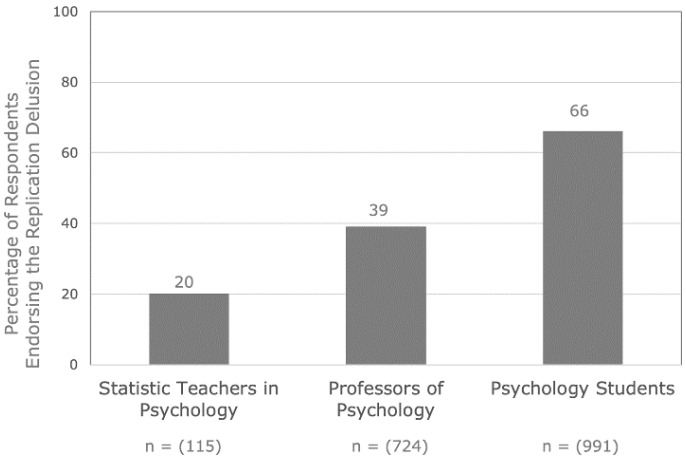

Figure 1. The replication delusion. Shown is the percentage of respondents who endorsed that $p = 0.01$ logically implies a 99% chance of replication. For details, see Gigerenzer (2018).

All in all, a total of 20% of the faculty teaching statistics in psychology departments considered the statement correct, as did 39% of professors of psychology and lecturers, and, unsurprisingly, two-thirds of 991 students.

The replication delusion is not alone in maintaining the belief in null hypothesis testing as a universal method. There is also the *illusion of certainty* (that significance disproves the null hypothesis and non-significance proves the experimental hypothesis) and *Bayesian wishful thinking* (that the *p*-value determines the probability of the null hypothesis or of the experimental hypothesis). In every study, the majority of researchers (56–97%) exhibited one or more delusions about what a significant *p*-value means (Gigerenzer, 2018).

All this leads to a positive conclusion: We should use the momentum of the replicability crisis to liberate research practice from methodological rituals and associated delusions, and we might conduct research to find out why we do what we do.

Financial support

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest

None.

References

Brunswik, E. (1956). Perception and the representative design of psychological experiments. Berkeley, CA: University of California Press.

Danziger, K. (1990). Constructing the subject: Historical origins of psychological research. Cambridge, UK: Cambridge University Press.

Fisher, R. A. (1935). The design of experiments. Oliver and Boyd.

Gigerenzer, G. (2006). What's in a sample? A manual for building cognitive theories. In Fiedler, K. & Juslin, P. (Eds.), Information sampling and adaptive cognition (pp. 239–260). New York: Cambridge University Press.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. Advances in Methods and Practices in Psychological Science, 1, 198–218. doi: 10.1177/2515245918771329.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. Psychological Review, 98, 506–528. doi: 10.1037/0033-295X.98.4.506.

Juslin, P., Winman, A., & Olssen, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. Psychological Review, 107, 384–396. doi: 10.1037/0033-295X.107.2.384.

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 754–770. doi: 10.1037/0278-7393.24.3.754.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5, 207–232.