<CT>**Replication is already mainstream: Lessons from Small-*N* designs**

<CA>Daniel R. Little and Philip L. Smith

<CAA>*Melbourne School of Psychological Sciences, The University of Melbourne, Parkville VIC 3010, Australia.*

daniel.little@unimelb.edu.au

philipls@unimelb.edu.au

http://psychologicalsciences.unimelb.edu.au/research/research-groups/knowlab

<C-AB>**Abstract:** Replication is already mainstream in areas of psychology that use small-*N* designs. Replication failures often result from weak theory, weak measurement, and weak control over error variance. These are hallmarks of phenomenon-based research with sparse data. Small-*N* designs, which focus on understanding processes, treat the individual rather than the experiment as the unit of replication and largely circumvent these problems.

<C-Text begins>

The claim that psychology has not given due consideration to replication treats psychology as a homogeneous discipline in which the focus is on demonstrating the presence or absence of experimental effects. In contrast, we argue that replication has long been a part of standard research practice, and is already mainstream, in several areas of psychology including visual and auditory psychophysics, animal learning,

mathematical psychology, and many parts of cognitive psychology. A common feature of research in these areas is the systematic use of small-*N* designs, in which a small number of expert participants (or highly-trained animals) are tested over many sessions. The effects of interest in these designs are thus replicated over trials, over sessions, and between participants. As a result, the questions of theoretical interest can be tested at the individual participant level rather than the group level. The individual participant rather than the group then becomes the replication unit, and the study effectively becomes its own replication.

Failure to replicate is often a symptom of deeper problems that arise from the three vices of weak measurement, weak theory, and weak control over error variance. It is typical in much of psychology for the relationship between the measurement scale and the underlying theoretical constructs to be at best ordinal. Ordinal-level theories can at best predict that performance in one condition will be greater (more accurate, faster, etc.) than performance in a second condition; they cannot predict strong functional relationships. They also tend to be sparse in the sense that inferences are made using single point estimates. Because effects vary from individual to individual, the typical response to measurement variability is to increase the sample size. Without addressing these more basic problems, however, a focus on increased replication will only squander limited resources on ill-thought out questions. Although often discussed in different terms, replication can be viewed simply as another way to increase the sample size to try to obtain a better estimate of the effect size. When viewed in this way, replication

continues to serve the questionable goal of establishing the existence of an effect that is defined only in weak ordinal terms.

In contrast to this type of phenomenon-driven research (Meehl 1967; 1990), the goal of small-*N* research is usually not to demonstrate some effect but to elucidate the underlying process-based mechanism that leads to the behavior of interest. This typically entails strong measurement and hypothesizing on a stronger-than-ordinal scale. In visual psychophysics, for example, variables such as contrast, summation time, motion direction or speed, or orientation thresholds or response time are measured on ratio scales and used to define functional predictions across the range of stimulation (e.g., psychometric functions or response time distributions).

The focus of process-based research is on testing theoretical model predictions and not on testing the significance of experimental effects. Because the mechanisms of interest are typically defined at the individual level, it is most appropriate to test predictions about them at the same level (Grice et al. 2017). To appropriately control error variance across individuals, at least two methods are commonplace: first, participants are typically highly-practiced; second, stimulus manipulations are tailored to the specific sensitivities of the individuals. These methods act to counteract the lack of precise control that arises with naïve participants. Individuals are tested extensively so that the distribution of responses (and other characteristics of those responses, like timing) is estimated with high power. Testing a small number of participants, each of whom acts as a replication of the entire experiment, controls for contextual variation

across both time (between sessions but within individuals) and individuals (between individuals but within sessions).

The upshot of this style of research is that rich contact can be made between theory and data. This contact facilitates the use of strong inference methods to falsify specific models (see e.g., Little et al. 2017) and testing of strong out-of-sample and out-of-context predictions (Yarkoni & Westfall 2017). These kinds of systematic tests of strong quantitative relationships are characteristic of mature sciences that psychology should be striving to become. While replication might weed out spurious effects, it often begs the question why we should care about these effects in the first place.

We (Smith & Little 2018) recently demonstrated the advantages of individual-level analysis in cognitive settings by simulating effects of different sizes using the additive factors method (a method for characterizing the stages of processing in a cognitive task; Sternberg 1969), and then estimating the power of either individual level analysis (e.g., maximum likelihood model estimation) or group-level statistical analysis (i.e., analysis of variance [ANOVA]). The goal of the additive factors method is to determine the presence or absence of an interaction which provides either falsification or confirmation, respectively, of the point prediction of a serial, sequential-stages processing model. Our results showed that the individual-level analysis could detect the presence of an interaction even with small effect sizes. The group-level statistical analysis, by contrast, only reached similar levels of power when the group sample size was increased

substantially. Further, the individual-level analysis also provides an estimate of the value of the interaction and the consistency with which it appeared across individuals.

Small-*N* designs will not be appropriate for all areas of psychology. They will not be appropriate with reactive measures that allow only a single measurement per person or when multiple measurements are made on individuals but the resulting data are sparse. In the latter eventuality, the best approach is to model the individual variation at the group level (i.e., hierarchically; Lee & Wagenmakers 2005). Our argument is that the level of replication must be appropriate for the question being asked. In the areas of psychology that we are concerned with, this is at the individual level. The fact that areas that routinely use small-*N* paradigms have so far remained immune to the replication crisis afflicting other areas of psychology can be seen as an object lesson on the kind of methodological reform that the discipline requires, which goes deeper than just the routine practice of replication.

<C-Text ends>

<RFT>

References[Daniel R. Little and Philip L. Smith][DRL]

<refs>

Grice, J., Barrett, P., Cota, L., Felix, C., Taylor, Z., Garner, S., Medellin, E. & Vest, A. (2017) Four Bad Habits of Modern Psychologists. *Behavioral Sciences* 7:53. [DRL]

Lee, M. D. & Wagenmakers, E. J. (2005) Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review* 112:662-68. [DRL]

Little, D. R., Altieri, N., Fific, M. & Yang, C-T. (2017) *Systems factorial technology: A theory driven methodology for the identification of perceptual and cognitive mechanisms.* Academic. [DRL]

Meehl, P. E. (1967) Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34:103-115. [DRL]

Meehl, P. E. (1990) Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports* 66:195-244. [DRL]

Smith, P. L. & Little, D. R. (2018) Small is beautiful: In defence of the small-N design. *Psychonomic Bulletin & Review*. [In Press]. [DRL]

Sternberg, S. (1969) The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica* 30:276-315. [DRL]

Yarkoni, T. & Westfall, J. (2017) Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science* 1-23. [DRL]