Looking to solve the replication crisis in psychology? Limitations of questionnaire methods must be considered.

Throughout its history, psychology has been faced with fundamental crises that all revolve around its disciplinary rigour. Current debates – led in Nature, Science and high-ranking psychology journals – are geared towards the frequent lack of replicability of many psychological findings. New research led by **Jana Uher** *highlights methodological limitations of the widely used questionnaire methods. These limitations may be a major source for the lack of repeatability in psychology.*

In previous debates, the causes of psychology's replication problem were sought in the technical details of the statistical methods of data analysis used, the samples studied, the lack of transparency in research practices or the publication biases. But in my recent research, I have highlighted a more fundamental cause not yet widely considered—the methods used in psychology research to generate the data in the first place.

In psychology and the social sciences more generally, standardised assessment methods such as rating scales are widely used for generating quantitative data. But despite their popularity, these methods contain serious methodological limitations that inevitably compromise the replicability of findings.

In recent research, funded by the German Science Foundation (DFG) and published in the April issue of the *Journal of Research on Personality*, I pinpointed essential methodological differences between assessments and observations by exploring the elementary question: What demands do the different methods place on the persons who generate the data?

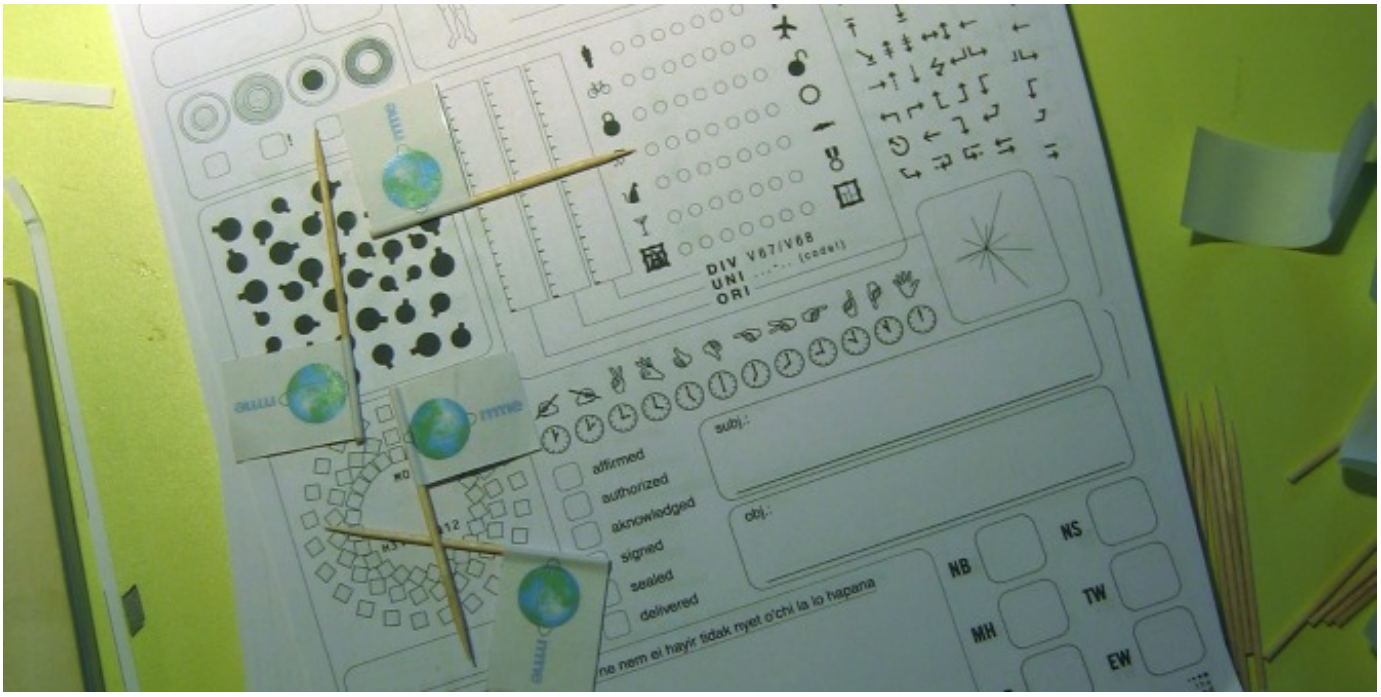

Image credit: up wanders – originally posted to Flickr as questionnaire (CC BY 2.0)

Behaviours are public phenomena; therefore, the same act can be perceived by multiple observers. This allows for training observers to categorise behaviours in the same ways and for establishing standardised rules of how to encode the events observed into data. Challenges arise because behaviours can be observed only in the moments in which they occur—in other words, real-time. But once observers have taken down the just-observed, their job in generating the data is completed.

Assessors, by contrast, must accomplish an entirely different task. To judge someone, raters must draw on the experiences that they have made in the past and must compare individuals with other individuals and over time. Thus, ratings are inherently memory-based and retrospective. But memory recall is known to be fallible and biased in many ways.

Further differences occur in the ways in which the data are encoded and the schemes that are therefore used. Observational categories are explicitly defined and observers are trained to understand and use them in the same ways. By contrast, the questions and answer categories of questionnaires are worded in everyday language and commonly no training occurs. Instead, raters must use their common-sense knowledge to interpret the meaning of the categories used for generating the data. But everyday concepts are fuzzy and context sensitive. What people may specifically have in mind when filling in a given questionnaire and how they assign their ideas to the categories marked on the scale is never enquired.

But the thought processes involved in assessments are highly complex. For example, to assess how “often” somebody tends to do something, raters must consider that some behaviours generally occur more often than others. Thus, to be judged as “often”, some behaviours must occur *more often* than other behaviours. This means that the *same* answer category encodes *different* information! Given this, what do the quantitative data derived from questionnaire markings actually signify?

I find it astonishing and alarming that even after a century of research in which questionnaire methods have been used intensely, we still do not know much about what is going on in people’s minds when they are generating assessments on a rating scale. Scientific quantification depends on the exact specification of what is to be measured and of how this is encoded into data. But questionnaire assessments do not fulfil these basic requirements. With such methods, how could replicability ever be achieved?

To illustrate these methodological limitations, I have conducted a five-method study in collaboration with Elisabetta Visalberghi from the Institute of Cognitive Sciences and Technologies, National Research Council of Italy (ISTC-CNR) in Rome. The study showed that, compared to observations, assessments contained numerous biases derived from raters’ stereotypical beliefs about individuals.

It also showed that raters’ interpretations of the standardised survey questions varied considerably because they considered diverse contexts that viewed the same question in a different light. As a consequence, raters’ interpretations overlapped with those intended by the researchers to only 54-77%! Thus, contrary to beliefs widespread in psychometrics and the social sciences, standardised questionnaire items do not reflect standardised meanings. As a consequence, standardised rating items do not enable repeated measurements of the same idea.

These profound methodological differences between assessments and observations have not been previously well considered. They offer a novel approach to explaining the frequent lack of replicability of findings in psychology.

This piece is based on a recently published journal article: *Uher, J., & Visalberghi, E. (2016). Observations versus assessments of personality: A five-method multi-species study reveals numerous biases in ratings and methodological limitations of standardised assessments. Journal of Research in Personality, 61, 61-79. http://www.sciencedirect.com/science/article/pii/S0092656616300058*

Note: This article gives the views of the authors, and not the position of the LSE Impact blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Jana Uher is a Senior Research Fellow and Marie Curie Fellow at the London School of Economics and Political Science. She has developed the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm), a novel paradigm aimed at making explicit and scrutinising the philosophical, metatheoretical and methodological foundations of psychological, behavioural and social-science research on individuals. Jana publishes widely on topics in theoretical psychology, philosophy of science, methodology, behavioural research, biology, primatology, evolutionary research, personality psychology, social psychology and cognitive sciences. @ResIndividuals