Generalizability, Replicability, and New Insights from Understudied Populations: Introduction to the Special Issue "Advancing Methods for Psychological Assessment Across Borders and Populations" in EJPA

Peter Adriaan Edelsbrunner[1*†]

Kai Ruggeri[2*†]

Kaja Damnjanović[3]

Samuel Greiff[4]

Jeremy Lemoine[5,6]

Matthias Ziegler[7]

[1]Department of Humanities, Political and Social Sciences, ETH Zurich, Switzerland.

[2] Department of Health Policy and Management, Columbia University, USA.

[3]University of Belgrade, Faculty of Philosophy, Department of Psychology, Laboratory for Experimental Psychology, Institute Of Philosophy, Serbia.

[4]University of Luxembourg, Luxembourg.

[5]Department of Psychology, University of East London, UK.

[6]ESCP Business School, UK.

[7]Department of Psychology, Humboldt-Universät zu Berlin, Germany

*Corresponding Author: Peter Edelsbrunner, peter.edelsbrunner@ifv.gess.ethz.ch, postal address: ETH Zurich, RZ H16, Claussiusstrasse 59, 8092 Zurich, Switzerland.

†PAE and KR shared the role as head guest editors for the special issue related to this editorial and are shared first authors.

Generalizability, Replicability, and New Insights from Understudied Populations: Introduction to the Special Issue "Advancing Methods for Psychological Assessment Across Borders and Populations"

Research on assessment that takes an individual differences-perspective is by definition dependent on the individuals and the specific populations that it studies. Consequently, if we study individuals from the same or similar underlying populations over and over again, we will miss important insights on the stability of findings, and how psychological traits are structured and play out in individuals with different characteristics and across different populations. This not only limits results to the population investigated, but also involves a high possibility for type 1 errors in overgeneralizing findings and their implications beyond the psychological and empirical rationale.

There are more and more initiatives from regions across the globe that aim to deal with the challenges posed by the lack of generalizability and replicability of psychological findings (Bishop, 2019; Earp & Trafimow, 2015). These include international or national networks of researchers aiming to promote research replicability (e.g., Center for Open Science, UK Reproducibility Network), encouragement to pre-register research studies (Eich, 2013; Van't Veer & Giner-Sorolla, 2016), research studies aiming to replicate previous findings (e.g., Visser et al., 2022), and free online platforms to share study protocols and data (e.g., Open Science Framework). In line with these initiatives, the motivation behind this special issue in EJPA (Ruggeri et al., 2019) is to encompass contributions that not only study sometimes neglected populations, but also combine this aspect with the approach of registered reports (Greiff & Allen, 2018). In this way, we intended to gather studies that tackle issues of generalizability, but also of replicability, within the same research designs.

As has been pointed out repeatedly and from different perspectives, generalizability and replicability are two interwoven issues, and this also applies to the field of psychological assessment. Fiedler (2011) described that not only samples, but also all other aspects of studies that might induce variation in results are commonly picked in ways to maximize observed effects. Whereas this allows researchers to better pin down theoretically expected effects, it can undermine the generalizability and replicability of observed effects. Barr et al. (2013) also pointed this out in the context of statistical modeling. They argued that in order to maximize the generalizability and replicability of effects, researchers should specify random variation in their statistical models across all factors that might induce variability by design.

Their arguments, however, do not only have statistical implications; they also point towards the theoretical consideration of often-neglected factors in study design and sampling. Specifically, if we want to generalize results across all design meaningful design factors in our studies, then random effects represent a statistical means for establishing generalizability. Similarly, Yarkoni (2017) argued for the specification of random effects for all variables that could induce theoretically meaningful variation. Approaches in which many analysts model the same data have also shown how researcher degrees of freedom in analysis can contribute to variation in findings and conclusions on the very same data (e.g., Schweinsberg et al., 2018; Hoogeveen et al., 2022).

In the present issue we aimed to present contributions that would tackle a lack of generalizability and replicability of findings due to a lack of variation in the assessed populations, as well as in researchers' degrees of freedom in conducting their studies and analyzing their data. This was achieved by assessing specific populations that have not been the focus of the respective research topics, and prioritizing studies that were pre-registered reports.

Trialing, extending, and comparing studies and their findings within and between populations that are seldom in the focus of research can help us question long-standing beliefs regarding untested assumptions and the generalizability of findings (e.g., Terraciano et al., 2005). Preregistration can contribute to factors of the theoretical and methodological quality of research which in turn might improve the reliability and replicability of results (Soderberg et al., 2021). With this foci, this special issue introduces a set of registered reports across various topics of psychological assessment to the readership of the journal. The hope behind this is that readers and (future) authors will see registered reports as a valuable tool that increasingly finds its way into the toolbox of research on psychological assessment.

How the Present Papers Contribute to Open Science and Shed Light on Generalizability and Robustness

In addition to assessing specifically selected populations and doing so mostly on the basis of registered reports, the set of contributions found in this special issue also tackled theoretical questions that are of high relevance to the assessment community. In this section, we summarize what the five contributions did and how, on top of the specific research question, they might contribute to our understanding of generalizability and replicability.

Twomey and Johnson (this issue) created norm tables of the UK and Ireland for the International Personality Item Pool (IPIP; Goldberg et al., 2006) NEO-300. They selected UK and Irish participants (N = 18,591) from a global IPIP-NEO-300 dataset within Johnson's IPIP-NEO data repository and identified the norms for different personality facets and different age groups. This allows researchers investigating personality using the IPIP-NEO-300 in the UK and Ireland to have norm tables to interpret the results of their own participants. Although this article

is not based on a registered report (unlike the other four articles published in this issue), this paper was included as its authors make the convincing case of providing norm tables for the United Kingdom and Ireland that were assessed using an open source assessment instrument (i.e., IPIP-NEO-300). These norm tables are valuable resources for researchers who have financial constraints and cannot (or would rather not) use proprietary measures of the Big Five. The study, with the inclusion of freely accessible and modifiable norm tables, contributes to an approach to open science that integrates the needs of researchers with little financial resources. Therefore, we believe that the study is fully in line with the scope and aims of this special issue.

Rachev et al. (this issue) examine whether the risky-choice and attribute framing effect, describing variation in preference under different presentations of the same problem (Tvesky & Kahneman, 1981), occurs and shows similar relations to the willingness and ability to think in line with rational norms across two samples. The willingness and ability to think in line with rational norms was represented by two constructs; namely, actively open-minded thinking (Haran et al., 2013) and bullshit receptivity (e.g., Ilić & Damnjanović, 2021). The authors found that the susceptibility to framing was associated with these two constructs in both samples, with a stronger negative relation of susceptibility to framing with actively open-minded thinking in Bulgaria than in North-America. Rachev et al.'s (this issue) paper contributes to the generalizability of findings in several ways. First, each construct was measured using multiple items or multiple tasks, and this allowed the authors to use multiple indicators to represent the construct at a broader level when using latent variable modeling. Second, the authors performed measurement invariance analyses; this enabled them to identify and correct for cross-cultural deviation in the measurement models and allowed them to perform cross-cultural comparison on an invariant model. Finally, this study exemplifies how pre-registered reports can incorporate

preregistered procedures for evaluating the fit of statistical models (cf. Greiff & Allen, 2018), which is of central relevance for assessment research in which latent variable models and their associated fit statistics are commonly implemented.

Clay et al. (this issue) investigate the relationship between willpower beliefs and perceived effort, dispositional constructs (personality traits), and satisfaction with reward, a topic that may be interesting to the researchers in cognitive psychology, decision sciences, individual differences, and neurosciences in general. The main hypothesis in this study is that people holding limited willpower beliefs perceive cognitive tasks as more effortful and associated rewards as less satisfying relative to people who do not hold such beliefs. In 187 participants from North-America, multilevel models indicate support for this hypothesis. In addition, the authors find that participants with higher levels of need for cognition perceived presented tasks (N-back tasks) as less effortful, whereas there were rather small and comparably few relations with personality measures from the Big Five-domain. On the basis of a registered report, this study demonstrates how preregistration ensures the reliability of findings involving constructs such as mental depletion that have been under allegations such as of p-hacking in recent years (Friese et al., 2019). Whereas this study does not involve an experimental manipulation of mental depletion, it shows that under preregistration, which is supposed to limit p-hacking and similar questionable research practices, self-ratings related to this construct can produce clear patterns of correlations.

Žeželj et al. (this issue) present the development and initial validation of a scale for assessing proneness to doublethink, which they define as the general proneness to tolerate inconsistencies within one's own beliefs or knowledge. Within a registered report, the authors find that the new scale on which they report approximates a one-factor structure and among other

results shows mostly consistent relations with conspiracist mentality, beliefs in conspiracy theories, and rational and intuitive thinking styles, across two samples. Similar to Rachev et al. (this issue), this study exemplifies how preregistration of model fit evaluation contributes to our robust understanding of phenomena and their replicability across samples. In their measure of belief in conspiracy theories, the authors included items specific to the Serbian context. They found stronger relationships between the concept of doublethink and their measure of belief in conspiracy theories specific to the Serbian context than between the concept of doublethink and a more general measure of conspiracist thinking. This result points to the relevance of including culture-specific items that are tailored towards non-WEIRD samples.

In the last contribution, Buabang et al. (this issue) validate a scale measuring perceived financial wellbeing across samples from six European countries and the US. The authors find that the scale, in contrast to their preregistered hypotheses, was not unidimensional but rather showed a two-dimensional structure of affective and behavioral financial stress. The authors further find that measurement invariance was not given across all countries and that only partial invariances could be achieved. Consequently, country comparisons using this measure might be undertaken based on models that correct for deviations, but the authors conclude that composite scores cannot be used for valid comparisons. This contribution demonstrates how preregistration allows reporting results that might appear suboptimal from a measurement perspective (meaning both instrument and analytical approach) but have been based on a-priori defined analyses, adding to their trustworthiness. Similar to the paper presented by Twomey and Johnson (this issue), this study highlights the importance of measurement invariance when conducting research across different populations (Vandenberg & Lance, 2000) as it allows for comparison across populations and greater generalization of the findings.

What do We Learn from these Contributions?

In acting as guest editors for this special issue in the European Journal of Psychological Assessment, we have learned a great deal about how cross-cultural research projects can contribute to our understanding of psychological assessment, and about the role that registered reports can play therein. As Rachev et al. (this issue) highlight with regard to the central construct in their study, it is sometimes implicitly assumed that psychological processes work independently of socio-cultural contexts. This is definitely not the case. Studies examining the actual empirical fit of data to such assumptions cannot only provide top-down tests of theories questioning generality; they can also provide a basis for the bottom-up-development of theories about the generalizability and boundary conditions of reliable and valid psychological assessment and the associated phenomena.

Finally, and perhaps especially noteworthy, most of the studies within this issue have been conducted by international teams of researchers from diverse countries. Many of the contributing researchers were still in the pre- or post-graduate phases of their studies during the active phases of their research. This context makes it even more challenging to engage into the long-term commitment that registered reports usually come along with. At the same time, this type of international research projects provide an optimal academic playground for establishing academic connections in which junior researchers can look beyond the established paradigms of their local research groups. Assessment experts at different institutions from different countries sometimes follow strongly diverging or even inconsistent approaches. An arguably prime example of this in psychological assessment are the diverging views that item response theory and Rasch modeling sometimes take on measurement and model fit (e.g., Edelsbrunner & Dablander, 2019). Junior academics who are educated at an institution specializing in Rasch

measurement theory will typically be conveyed views and assumptions that vary substantially from those conveyed at institutions with a history of research that is grounded in item response theory. This example showcases that, although psychological assessment is partially grounded in formal measurement theory, it is also a highly theoretical field in which incongruent perspectives can inform one another to contribute to theoretical and methodological progress. By gathering and presenting these contributions in this issue, we hope to inspire assessment-researchers to consider broadening their research networks, as well as those of their affiliate junior academics, beyond their typical borders of state, age, and identity, engaging in activities that promote and foster transparency and replicability collaboratively.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature, 568*(7753), 435. https://doi.org/10.1038/d41586-019-01307-2

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6*, 621. https://doi.org/10.3389/fpsyg.2015.00621

Edelsbrunner, P. A., & Dablander, F. (2019). The Psychometric Modeling of Scientific Reasoning: A Review and Recommendations for Future Avenues. Educational Psychology Review. https://doi.org/10.1007/s10648-018-9455-5

Eich, E. (2013). Business Not as Usual. *Psychological Science, 25*(1), 3-6. https://doi.org/10.1177/0956797613512465

Fiedler, K. (2011). Voodoo correlations are everywhere—Not only in neuroscience. *Perspectives on Psychological Science*, *6*(2), 163–171. https://doi.org/10.1177/1745691611400237

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2019). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review, 23*(2), 107-131. https://doi.org/10.1177/1088868318762183

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality, 40(1), 84-96. https://doi.org/10.1016/j.jrp.2005.08.007

Greiff, S. & Allen, M. S. (2018). EJPA introduces registered reports as new submission format. *European Journal of Psychological Assessment, 34,* 217-219. https://doi.org/10.1027/1015-5759/a000492

Ilić, S., & Damnjanović, K. (2021). The effect of source credibility on bullshit receptivity. *Applied Cognitive Psychology, 35*(5), 1193-1205. https://doi.org/10.1002/acp.3852

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E.-J. (2022). *A Many-Analysts Approach to the Relation Between Religiosity and Well-being* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/pbfye

Maraun, M., & Heene, M. (2015). A note on the implications of factorial invariance for common factor variable equivalence. Communications in Statistics - Theory and Methods, 00–00. https://doi.org/10.1080/03610926.2014.917186

Ruggeri, K., Damnjanović, K., Edelsbrunner, P. A., Lemoine, J. E., & Ziegler, M. (2019). Call for Papers “Advancing the Reproducibility of Psychological Assessment Across Borders and

Populations". *European Journal of Psychological Assessment*, *35 (2)*, 295-296. https://doi.org/10.1027/1015-5759/a000523

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., van Assen, M. A. L. M., Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S. M. G., Robinson, D., … Luis Uhlmann, E. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249. https://doi.org/10.1016/j.obhdp.2021.02.003

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., ... & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour, 5*(8), 990-997. https://doi.org/10.1038/s41562-021-01142-4

Terracciano, A., Abdel-Khalek, A. M., Adam, N., Adamovová, L., Ahn, C. K., Ahn, H. N., ... & McCrae, R. R. (2005). National character does not reflect mean personality trait levels in 49 cultures. *Science, 310*(5745), 96-100. https://doi.org/10.1126/science.1117199

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458. https://doi.org/10.1126/science.7455683

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. Journal of Experimental Social Psychology, 67, 2-12. https://doi.org/10.1016/j.jesp.2016.03.004

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177/109442810031002

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., ... & Zettersten, M. (2022). Improving the generalizability of infant psychological research: The ManyBabies model. Behavioral and Brain Sciences, 45. I: https://doi.org/10.1017/S0140525X21000455

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685