Pre-print, version 2 (January 2023).

Please contact the first author for the final version before citing.

Generalizability Crisis Meets Heterogeneity Revolution: Determining Under Which Boundary Conditions Findings Replicate and Generalize

Julia Moeller[1], Julia Dietrich[2], Andreas B. Neubauer[3], Annette Brose[4], Jana Kühnel[5], Mathias Dehne[1], Miriam F. Jähne[2], Florian Schmiedek[3], Henrik Bellhäuser[6], Lars-Erik Malmberg[7], Kristina Stockinger[8], Michaela Riediger[2], and Reinhard Pekrun[9 10]

[1] Leipzig University, Germany
[2] Friedrich Schiller University Jena, Germany
[3] DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany
[4] Humboldt University of Berlin, Germany
[5] University of Vienna, Austria
[6] University of Mainz, Germany
[7] University of Oxford, United Kingdom
[8] University of Augsburg, Germany
[9] University of Essex, United Kingdom
[10] Australian Catholic University, Sydney, Australia

Corresponding author: Julia Moeller, Leipzig University, Marschnerstr. 31, 04109 Leipzig, Germany, E-mail: julia.moeller@uni-leipzig.de

Acknowledgements: We would like to thank Julia M. Rohrer, Laura Bringmann and Peter Kuppens for helpful and thought-provoking comments on the manuscript. This work was supported by a Jacobs Foundation Early Career Research Fellowship to Julia Moeller and a grant by the German Research Foundation (DFG; #451682742) to Julia Moeller and Julia Dietrich.

Abstract

Intensive longitudinal studies typically examine phenomena that vary across time, individuals, contexts, and other boundary conditions. This poses challenges to the conceptualization and identification of replicability and generalizability, which refer to the invariance of research findings across samples and contexts as crucial criteria for trustworthiness. Some of these challenges are specific to intensive longitudinal studies, others are similarly relevant for the work with other complex datasets that contain multilayered sources of variation (individuals nested in different types of activities or organizations, regions, countries, etc.).

This article opens with discussing the reasons why research findings may fail to replicate. We then analyze reasons why research findings may falsely appear to be non-replicable when in fact they were as such replicable, but lacked generalizability due to heterogeneity between samples, subgroups, individuals, time points, and contexts. Following that, we propose conceptual and methodological approaches to better disentangle non-replicability from non-generalizability and to better understand the exact causes of either problem. In particular, we apply Lakatos's proposition to examine not only whether but under what boundary conditions a theory is a useful description of the world, to the question whether and under which conditions a research finding is replicable and generalizable. Not only will that contribute to a more systematic understanding of and research on replicability and generalizability in longitudinal studies and beyond, but it will also be a contribution to what has been called the heterogeneity revolution (Bryan et al., 2021; Moeller, 2021).

Keywords: intensive longitudinal studies, replicability, generalizability crisis, heterogeneity revolution, boundary conditions, Lakatos

Generalizability Crisis Meets Heterogeneity Revolution: Determining Under Which Boundary Conditions Findings Replicate and Generalize

"(...) intellectual honesty consists rather in specifying precisely the conditions under which one is willing to give up one's position."

Lakatos (1978, p. 9)

"This [heterogeneity] revolution will be defined by the recognition that most treatment effects are heterogeneous, so the variation in effect estimates across studies that defines the replication crisis is to be expected as long as heterogeneous effects are studied without a systematic approach to sampling and moderation."

Bryan et al. (2021, p. 1)

People reading about scientific findings expect them to be trustworthy, usually expecting that another researcher would draw similar conclusions if studying the same phenomenon with the same methods, i.e., replicate the findings. Failed replication attempts raise questions about the trustworthiness of either the original study's finding, the replication attempt, or both. Recently, the replication crisis revolved around the insight that many research findings were not replicated in independent studies (Ioannidis, 2005; 2012; Nosek et al., 2022). The ensuing search for the causes of the limited trustworthiness of research findings led to debates about flawed research methods and questionable research practices (for overviews, see Appendix A; Bakker et al., 2012; Lundh, 2019; Syed, 2021).

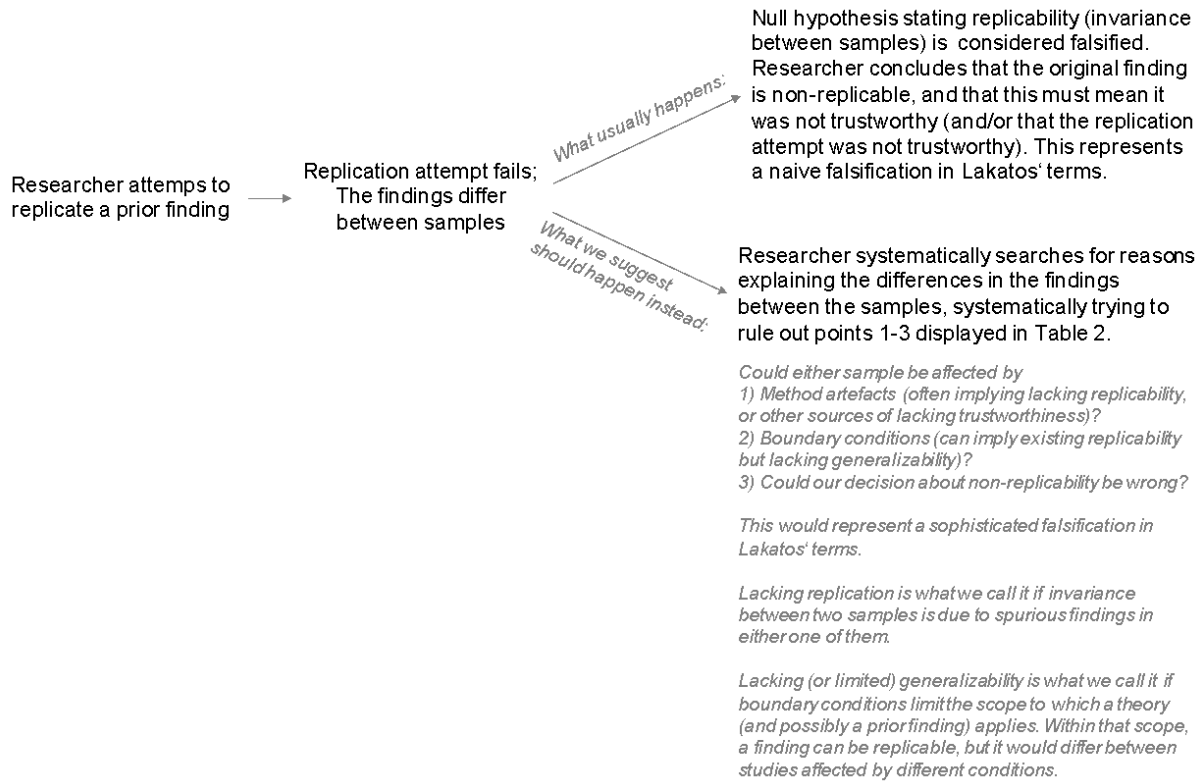
In this article, we discuss further reasons that can limit the replicability and generalizability of research findings, beyond the much-discussed flawed research methods mentioned above. In particular, we address why it can occasionally be misleading to judge a research finding as not trustworthy if it differs between an original sample and a replication

attempt in a novel sample. We discuss how heterogeneity between individuals, time points, and other context characteristics can limit the generalizability of a research finding and why it is logically impossible to distinguish non-replicability from non-generalizability with certainty (known as the hidden moderator problem). To help mitigating the practical implications of this logical problem, we discuss an epistemological framework helping researchers to make at least educated guesses about the exact reasons of why a given research finding may differ between various studies, samples and populations. The goal is to help researchers disentangle as well as possible whether a research finding differed between samples due to

1) method artifacts implying lacking trustworthiness of the finding in the original sample and/or the replication attempt,
2) person-, time- or context-specific boundary conditions limiting the generalizability, but not other aspects of trustworthiness (including replicability, repeatability robustness) of a research finding, or
3) logical fallacies, lacking or flawed decision-making criteria and other reasons limiting the objectivity of researchers when making inference about replicability and generalizability (e.g., confusing the above-mentioned points 1 and 2, or lacking clear criteria and methods to determine whether two findings are sufficiently similar to be considered invariant across samples (for a more detailed explanation of each point, please see Table 2).

In this article, we propose that it only makes sense to treat the invariance of research findings across samples as a litmus test of the trustworthiness of research, if we can reasonably well distinguish between the above-mentioned factors affecting judgments about (in-)variance of research findings across different samples (see Figure 1).

Figure 1: This article in a nutshell:



We discuss all of these issues mentioned above from our perspective as researchers experienced in the work with intensive longitudinal data (ILS). ILS data are typically collected in real-life contexts, which include more multilayered, heterogeneous, and messy sources of variation than most controlled lab experiments that so far have been the main focus of research on replicability. We believe that encountering such uncontrolled and manifold sources of heterogeneity, along with the partially exploratory nature of many intensive longitudinal studies gives this article a novel perspective. This goes beyond the typical hypothetico-deductive perspective emphasized in most previous publications on replicability and generalizability. The implications of the insights gained in ILS for the broader understanding of generalizability in Social and Personality Psychology are discussed on page 26 below. We aim to connect current debates about generalizability and boundary conditions (e.g., Yarkoni, 2022a; Busse et al., 2017; Deffner et al., 2022) to other ongoing debates in Personality and Social Psychology research that help us understand the implications of

sources of heterogeneity for generalizability. These include the debate about person-oriented methods (e.g., Molenaar, 2004; Lundh, 2022), the debate about a heterogeneity revolution (e.g., Bryan et al., 2021) the debate about integrations of idiographic and nomothetic approaches (e.g., Beck & Jackson, 2020; Beltz et al., 2016). For a summary of these debates, please see Figure 2. Finally, we provide a reading list for readers interested in debates about generalizability in Appendix D and have compiled a list of specific steps that various authors have proposed to improve our understanding of generalizability in the presence of heterogeneity and boundary conditions (Appendix C).

The challenges ILS posed to replicability and generalizability have been discussed elsewhere (Moeller et al., in prep.). They include, among others, the often limited samples of individuals and contexts, the lack of validated measures, the many novel, unexpected and exploratory findings warranting replications, and a need for formalized theories describing situation- and context-specificity and heterogeneity across individuals, contexts, and time points. This present article focuses on the need to *rethink the concepts* of replicability and generalizability *and the epistemological foundations of how to study them* when examining phenomena that change over time, and/or are highly dependent on so far unknown contextual characteristics.

Defining replicability, generalizability, and other aspects of trustworthiness

Articles of replicability and generalizability currently use different definitions of these terms, leading to jingle fallacies (same construct being called different names) and jangle fallacies (different constructs being labeled the same; see Block, 1995). To avoid such terminological ambiguities, we specify our working definitions in the following. Much has been written about definitions and types of *replicability* (see e.g., Association for Computing Machinery, 2020; Bollen et al., 2015; Feest, 2019; Goodman et al., 2016; Hardwicke et al., 2018; Schloss, 2018; Simons, 2014; Plesser, 2018; Plucker & Makel, 2021; Whitaker, 2017). Based on these previous discussions, we use the following framework to distinguish between

replicability, generalizability, and other aspects of the trustworthiness of a research finding. Since this article mainly focuses on replicability and generalizability, these terms are explained more in detail below Table 1. The other aspects of trustworthiness mentioned in the table are described in Appendix A.

Table 1: Overview of the terminology used in this article to characterize conditions under which findings of an original study remain invariant across investigations

	Similar boundary conditions (e.g., same culture, time, situation characteristics)			*Different boundary conditions, (e.g., different times, different types of individuals)*
	Same data		*Different data*	*Different data*
	Same research team	*Different research team*	*Same or different team*	*Same or different team*
Same methods aims: (to demonstrate invariance across researchers and samples)	Repeatability	Reproducibility	Direct replicability	Direct generalizability
Alternative methods, capturing same phenomena (aim: to rule out method artifacts)	Robustness	Robustness	Conceptual replicability	Conceptual generalizability

Note. This Table and its distinction between robustness, repeatability, replicability and generalizability was adopted and amended based on the Table 1 on page 3 in Schloss (2018); who bases his taxonomy on Whitaker (2017). We combined their definitions with the distinctions between same versus different data, same versus different research teams and same versus different methods proposed by the (Association for Computing Machinery, 2016). The distinction between direct and conceptual replicability was adopted from e.g., Feest, (2019) and Simons (2014), whereas the distinction between direct and conceptual generalizability is our suggestion.

Replicability is defined here as "the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected" (Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015, p. 6, see also Schloss, 2018; Whitaker, 2017). We distinguish between direct and conceptual replication (e.g., Feest, 2019; Simons, 2014).

Direct (also called exact) replication refers to invariance of results when the exact same sampling procedure, materials, methods, and analyses are used across data collections. *Conceptual replication* addresses to invariance of findings when the replication study used different stimuli, procedures, or analyses to capture the same phenomena or effects as the original study. The main purpose of direct replications is to rule out sampling biases and sample-related method artifacts (e.g., overfitting of prediction algorithms). The main purpose of conceptual replications is to rule out method artifacts (Hendrick, 1990; Stroebe & Strack, 2014), similar to the concept of robustness checks, with the difference that conceptual replications test *invariance between methods across samples* and robustness checks examine invariance between methods within a sample. To avoid confusion, we would like to point out that what other authors (and us) call direct or exact replication resembles what Goodman's 2016 describes as methods reproducibility combined with what the author describes as results reproducibility (for definitions of reproducibility, see Appendix A).

We define *generalizability* as finding similar results across populations (in line with definitions by Schloss, 2018; Whitaker, 2017) or across other boundary conditions, including person, context, and time characteristics. In line with the distinction between direct and conceptual replicability (see above), we propose a distinction between *direct generalizability* (finding similar results when the same methods are applied in new contexts) and *conceptual generalizability* (finding similar results with alternative methods in new contexts). Importantly, generalizability refers to the (absence of) sample- and population-specific boundary conditions of an original finding. A finding may be valid universally (which we call *global generalizability*), or only in a subgroup of individuals, or only in a limited set of contexts, countries, situations, time points, or only in the lab but not in everyday life situations (which we call *local generalizability,* see e.g., Moeller et al., 2022b).

Please note that different authors define generalizability in slightly different ways. For Yarkoni (2022a) and Flake Luong; & Shaw (2022), investigating generalizability includes making sure that a research conclusion is invariant across different operationalizations (e.g., different measures and research designs) that are supposed to capture the same phenomenon. Shadish, Cook, & Campbell (2002, P. 20) propose the more specific term *construct validity generalizations* for "inferences about the constructs that research operations represent". As we show in Table 1, we (along with many other authors) call it a matter of *robustness* if the same original dataset undergoes sensitivity / robustness analyses (e.g., comparing different measures, data collection procedures or analyses within the original sample). If operationalizations such as measurement instruments or research designs are compared between the original sample and a new sample to rule out method artifacts, we call that a conceptual replication, provided that no boundary conditions differ between the two samples. If the same comparison of method invariance is done for two samples that differ in relevant boundary conditions, we suggest to call that a conceptual generalization. If authors need an umbrella term for all three tests (robustness, conceptual replication and conceptual generalization), we suggest they call that testing for measurement invariance and mention explicitly that the aim is to rule out method artifacts, to avoid jingle and jangle fallacies.

Please note that some authors use the term external validity to refer to certain aspects of generalizability (e.g., Czibor et al., 2019; Pearl & Bareinboim, 2014). For instance, Shadish et al. (2002, p. 20; see also page 22) define external validity as "inferences about whether the causal relationship holds over variations in persons, settings, treatment, and measurement variables" and Czibor et al. (2019, p. 8) call generalizability "also known as external validity".

When studying context-specific phenomena, which is typical for ILS, a crucial question is whether a research finding is invariant in a new sample under different conditions,

such as different populations, time points, or contextual settings. Studying the generalizability of research findings aims to determine the scope of validity of the theory tested in the current study, whereas examining replicability asks whether research findings are independent from method artifacts, methodological flaws, and subjective researcher biases[1].

Below we discuss boundary conditions limiting the replicability and generalizability of research findings that are particularly salient in research with ILS. Although these points are relevant to solving the generalizability crisis (Yarkoni, 2022a), they have not yet been widely discussed in this context.

Conditions Limiting the Replicability and Generalizability

Replicability and generalizability have been conceptualized and tested primarily within the epistemological hypothetico-deductive framework (see Munafò et al., 2017), implementing Popper's principle of empirical falsification of *a priori* formulated hypotheses. Furthermore, suggestions for how to improve replicability following the wake of the replicability crisis have focused on strategies that nudge, encourage or force researchers to more consequently adhere to the hypothetico-deductive research approach (e.g., Munafò et al., 2017). Within that logic, we test the hypothesis that a previous finding is trustworthy and, therefore, replicable. Failing to replicate a previous result leads to the conclusion that the original result was not trustworthy, or that the current study's methods were flawed, or both. If further replication attempts confirm the non-replicability, then the theory predicting the non-replicable finding is typically considered falsified.

However, using the replicability of a previous finding as a criterion of a theory's trustworthiness has its limitations. As Popper (1935) mentioned and Lakatos (1978) elaborated long ago, it can be insightful to ask not only whether but *under which boundary conditions* a hypothesis is falsified. Lakatos argued that it might be productive not to stop

[1] If the methods applied to study a theory's scope of validity (i.e., generalizability) fail to capture heterogeneity, as we argue in this article, logically the attempt to study generalizability also becomes a question about method artifacts, similar to replicability, see Yarkoni (2022a).

when a hypothesis has been falsified but to test auxiliary hypotheses with comprehensive research programs to determine the exact conditions under which a theory is a useful model for describing the world. We propose to apply this approach, called ‘sophisticated methodological falsificationism‘ (Lakatos, 1978), to the hypothesis that a research finding is replicable and/or generalizable. This requires researchers to examine the various conditions that potentially limit the replicability and generalizability step by step to determine for each research finding under which exact circumstances it can be replicated and generalized. Lakatos originally referred to the scope of validity of *theories* by asking under which boundary conditions a *theory* is valid. In contrast, the debates following the replicability crisis taught us that we must first examine the scope of trustworthiness of our *research methods and research practices* by asking under which conditions they tell us anything about the trustworthiness of any research finding. Accordingly, Table 2 first summarizes conditions of replicability and generalizability related to research methods and practices (conditions 1.1-1.8), which have been broadly discussed since the replicability crisis and are therefore described more in detail in our Appendix B.

Table 2: *Conditions Limiting the Replicability and Generalizability of Research Findings*

1. Lacking trustworthiness of research findings due to flawed methods, flawed theoretical work, and flawed research practices

1.1 Flaws in theorizing, defining concepts, and hypothesis formulation
("theory crisis"; Eronen & Bringmann, 2021)

1.2 Lacking conceptual clarity about phenomena and estimands
(Bringmann et al., in press; Lundberg et al., 2021; "construct validity crisis": Schimmack, 2019)

1.3 Flaws in design and instruments of data collections
("measurement crisis"; Flake & Fried, 2020; "construct validity crisis": Schimmack, 2019)

1.4 Flaws in the sampling procedure (e.g., small samples leading to untrustworthy p-values and effect sizes, non-representative samples leading to biased findings)

1.5 Flaws in research analyses leading to untrustworthy or misleading empirical findings

1.6 Flaws in how empirical findings are reported
(Bakker & Wicherts, 2011)

1.7 Flaws in interpreting research findings
("inferential crisis": Starns et al., 2019; Syed, 2021)

1.8 Contributing to the above-mentioned issues: Flaws in the research infrastructure (incentive systems, publishing and funding decisions) leading to a lack of cumulative knowledge building
("normativity crisis", Lundh, 2019)

2. Boundary conditions causing discrepancies between the original study and the replication attempt, despite of either study being possibly trustworthy and replicable under invariant conditions, but not generalizable (i.e., not invariant across the conditions mentioned below)

2.1 *Person-specificity* (finding only true in certain samples of individuals)

2.2 *Time-specificity* (finding only true in certain time span)

2.3 *Context-specificity* (finding only true in certain locations & settings)

2.4 Statistical analyses failing to capture heterogeneity hiding behind overall trends
("validity crisis": Lundh, 2019; "generalizability crisis": Yarkoni, 2022a)

2.5 Other unknown or poorly understood boundary conditions (e.g., hidden moderator debate
Wingen et al., 2020)

3. **Logical fallacies, lacking decision making criteria and other sources of lacking objectivity of researchers making inference about replicability and generalizability**

3.1 Lack of clear, useable, and binding criteria to decide whether and to what extent findings were sufficiently similar to previously reported findings

3.2 Legitimate degrees of freedom in decisions about conducted analyses and inferences, multitude of alternative explanations, and of analytic and epistemological paths, leading to justified variation in conclusions drawn by different researchers.

3.3 Mistaking a lack of replicability for a lack of generalizability (attributing differences in findings between the original sample and a replication attempt to true non-replicability when they are due to unrecognized boundary conditions)

4. Unexplained non-replicability, i.e., the original finding did not withstand further testing and turned out to be spurious, which can happen even if state-of-the-art methods were properly applied & boundary conditions considered.

We propose that the conditions 2.1- 2.5 are particularly salient in research with ILS collected in real-life settings (see e.g., Moeller, Dietrich, & Baars, revise & resubmit). They have received less attention and are described below. They deal with uncontrolled and complex sources of variation that are common in ILS. Importantly, they point out that lacking generalizability due to heterogeneity between persons, time points, contexts, and other boundary conditions can be confused with a lack of replicability.

Who, Where, and When?: Findings may be Specific to Certain Conditions (but as such Trustworthy)

Person specificity[2] (2.1 in Table 2)

If individuals from one sample differ from another sample in factors (e.g., psychological or demographic characteristics) that affect the studied phenomenon, then it is possible that statistical coefficients obtained in either sample are trustworthy in the sense of being repeatable, robust, even theoretically replicable, but not generalizable. For instance, the first author attempted to replicate a previous finding, according to which school students show no gender differences in anxiety towards Math and Science lessons if experience sampling method -ESM- measures are used (trying to replicate some findings reported by Goetz et al., 2013). The original sample comprised of German students. The authors suggested that the findings reflected general features of experience sampling method measures. Therefore, there was no reason to assume that the country of origin would play a role or limit the finding to participants of certain nationalities. Our first author found the opportunity to test the hypothesis (no gender differences in situational ESM measures of math or science anxiety) in a sample of US high school students' feelings in school situations (Moeller et al., 2020), and

[2] We distinguish between sampled individuals, sampled time points, and sampled context characteristics (such as location or activity types). This point here refers to the person characteristics differing between samples and populations, whereas time points and context settings are addressed subsequently.

found to her surprise that females reported significantly higher anxiety than males. That raised the question: Was the original finding non-replicable, meaning possibly due to method artifacts? Or was it as such trustworthy and would have been replicable in a German sample, but failed to replicate due to systematic country differences? Further reflections suggest that the picture is more complex than that (differences in how, when and where the ESM surveys were gathered and inconsistent findings in further samples call for more systematic explorations of the boundary conditions).he example nevertheless illustrates the possibility that two findings may be valid (i.e., properly describe phenomena in the real world) and possibly replicable if research conditions are invariant across samples, but still differ between samples due to sample differences in person or context characteristics (which we call lacking generalizability).

Whether either of these sample-specific findings can be generalized to the overall population the tested theory aspired to describe is a matter of representativeness. For example, stress reactivity may differ largely depending on person characteristics present in a sample: Reduced reactivity is associated with age or culture (Stawski et al., 2019); enhanced reactivity is associated with mental health problems. If stress reactivity is studied in samples that differ in these conditions (age, culture, mental health) while researchers test nomothetic hypotheses and are unaware of age, culture or mental health acting as possible moderators, then invariance between their findings in the different samples may be mistaken for a lack of replicability when in fact it may represent a lack of generalizability (see Table 1 for the difference). It is therefore crucial to disentangle replicability from generalizability (see also Deffner et al., 2021; Simons et al., 2017).

Time specificity/non-stationarity[3] (2.2. in Table 2)

Over time, phenomena can emerge, disappear, or fluctuate. That means an original study may have described an effect accurately; the effect may then have changed and, therefore, differed from the first one in a subsequent replication study. Here, both results can be as such trustworthy. In this case, whether an effect looks the same in two studies is not a useful criterion for the trustworthiness of the research, which is similar to a low retest reliability being no useful criterion for the trustworthiness of a measurement instrument supposed to capture inherently fluctuating phenomena. Developmental dynamics thus require consideration in replication attempts, both at the theoretical and methodological level (e.g., Dietrich et al., in press). Some researchers might argue that an effect that has disappeared does not matter anymore. We would respond that such an argument would imply telling the population that the pandemic did not affect their well-being two years ago, because the effect has since disappeared since and disappeared effects cannot be theorized about. Not only Historians might have a hard time accepting such reasoning. If we follow that argument, we would have to tell a patient that you do not believe that last year they harmed themselves when were depressed, because this year (after some therapy) we see no correlation between their depression levels and self-harm probability, and since the relation between depression levels and self-harm has disappeared, it never happened for us researcher, nor do we think it should be taken as a warning sign for the future. We would miss the opportunity to take preventive therapeutic measures in case depression levels and lack of coping capacities might ever spiral out of control again. Fortunately, non-stationarity is increasingly examined (e.g., Casini et al., 2020), not least in the research on psychological disorders.

[3] Non-stationarity: Estimates (e.g., correlation coefficients, structure models) changing over time (e.g., Bringmann et al., 2022).

Context[4]-specificity (2.3 in Table 2)

Since findings can be specific to certain settings and locations (e.g., work versus leisure, urban versus rural), the invariance, i.e., generalizability, of measurement, structural and process models across contextual settings need to be tested empirically (Yarkoni, 2022a). This is particularly relevant in studies examining phenomena that are theoretically expected to differ between contexts, including most experience sampling method studies, which are often designed to capture context-specific phenomena (Larson & Csikszentmihalyi, 2014). Paradoxically, many experience sampling method studies are rather limited in their scope of contexts they capture (e.g., collecting data in only a few workplaces or schools; Lathia, Rachuri, Mascolo, & Rentfrow, 2013), in part due to such studies' high costs and efforts. Notwithstanding, the findings are often interpreted as being universally valid (interpreted as general laws of human behavior), although we cannot rule out that they are specific to the contextual settings (e.g., types of activities) where the data was gathered.

It is a matter of definition whether differences between cultures and countries should be defined as examples of context-specificity (context refers to locations and settings, as do countries) or person-specificity (since culture is not a location but a characteristic of groups of individuals). Ideally, both aspects should be disentangled. In any case, individuals in different populations may function differently. Therefore, many coefficients (averages, variances, the form of distributions, measurement properties, correlations, etc.) can vary between different countries and cultures (Georgas et al., 2004; Hofstede, 2001). Cultural biases in psychological studies' samples imply a limited knowledge about their findings' generalizability (e.g., Bryan et al., 2021). As Henrich et al. (2010) pointed out, most samples in psychological studies are

[4] To distinguish between time and space, we define context here as location, setting, circumstance, excluding the time-specificity addressed in point 2.2 in Table 2.

collected in western, industrialized, rich, educated and democratic countries, see Henrich et al., 2010). The problem is that researchers already ask limited questions about possible contextual boundary conditions, such as country-specific, culture-specific, economy-specific, or policy-specific factors. Then they collect such limited samples, leading to systematic bias and reduced information about context heterogeneity in the collected data (Henrich et al., 2010). Then they use methods that are unable to detect heterogeneity and individuals that function different than group trends (e.g.,averages) suggest, even if such heterogeneity was originally captured by the sampling and measurement procedures (e.g., Molenaar, 2004; Moeller, 2021). Then they get away with very limited considerations of boundary conditions in their publications' discussion sections, often being discouraged from speculating about possible boundary conditions that were not studied and demonstrated in the study at hand. Together, this creates a systematic blindness towards heterogeneity and boundary conditions (see Figure 2).

Statistical Analyses Failing to Capture Heterogeneity Hiding Behind Overall Trends (2.4 in Table 2)

If observations are nested within groups (e.g., repeated measures nested in individuals nested in work groups nested in organizations), then coefficients (e.g., averages, correlations, distributions) can differ between levels of analyses (here: situation-level, person-level, work group level, organization level). This is called lacking ergodicity (Molenaar, 2004), Simpson's paradox (Simpson, 1951) or non-homology (Klein & Koslowski, 2000). In ILS, this implies that correlations among variables measured repeatedly per person, including the factor structure, may differ if being calculated within versus between individuals (Brose et al., 2015; Kievit et al., 2017; Ram et al., 2013; Schmiedek et al., 2020, Schmitz, 2006). This poses challenges to inferences about replicability if a replication attempt unwittingly

estimates a coefficient on a different level than the original study (e.g., examining between-person variability versus within-person variability). Lacking ergodicity implies that research findings and conclusions obtained with the commonly used between-person analyses cannot be generalized to the individuals in a sample or population (McManus et al., under review; Moeller, 2021). A partial solution makes sure that coefficients that shall be compared across samples are calculated on the same level of analysis (Voelkle et al., 2014). Another solution proposed an ergodicity index informing researchers about lacking ergodicity (Golino et al., 2022). Some levels of analysis are considered habitually, such as students (level 1) nested in classrooms or schools (level 2) in educational research (Frenzel et al., 2007; Pekrun et al., 2019), and time points (level 1) nested in individuals (level 2) in ILS (Murayama et al., 2017). However, other levels of analysis might be overlooked. For example, if two studies on stress reactivity were conducted in the same city but at different universities, findings may differ because of, all else being equal, participants belonging to different neighborhoods. Unless hidden moderators/levels of analysis with influence on parameter estimates can be ruled out, heterogeneity in findings across studies might be due to unrecognized Simpson's paradox/lacking ergodicity.

When establishing the replicability, we typically compare sample-level (group-based) coefficients and expect them to represent general laws (e.g., average treatment effects, or the one correlation coefficient for the sample). This follows the nomothetic rationale expecting all (or most) individuals to function according to one and the same law (e.g., Hamaker, 2012; for. Review, see Robinson, 2021), which has been criticized as often unrealistic and of limited use (Beck & Jackson, 2020; Bolger et al., 2019; Bryan et al., 2021; Moeller, 2021; Molenaar, 2004; Richters, 2021). Richters (2021) argues that we neither know whether all individuals are affected by the same causal mechanisms, nor whether the path of causal transmission

(direct versus mediated, bidirectional, etc.) is the same across all individuals. It could be argued that a more sophisticated research on causal relations might mitigate that problem (see e.g., Deffner et al., 2021, elaborating on Pearl, 2018). A problem limiting the scope of that partial solution is that such research on causality still largely relies on the analysis on between-person variation and group trends of central tendency (e.g., averages, average treatment effects, sample-level correlation or regression coefficients). Such between-person coefficients are often not generalizable to some or even all individuals in many samples (see Figure 2 and the debates about ergodicity and person-oriented methods, e.g., Moeller, 2021; building on Molenaar, 2004). Some authors, however, have combined causal inference with within-person methods examining heterogeneity in causal processes, see Bolger et al. (2019).

If we cannot be sure for our sample-level coefficient (e.g., average treatment effect) to represent a general law or to describe the individuals in our sample well, what is our logical rationale to expect a second average in a second sample to represent the same law that the first sample-level coefficient already might not represent? It is a gap in the logic underlying comparison of averages across studies.

Furthermore, frequently used statistical coefficients may be ill-suited to describe sample-level trends in heterogeneous samples: Some statisticians suggest that in bimodal or other mixture distributions, or distributions with outliers, the average is no useful indicator of the central tendency (Derrible & Ahmad, 2015; Wirtz & Nachtigall, 1998). A multimodal distribution can imply that correlation coefficients do not represent the relationship between two variables in the way we think they do on any level of analysis (Matejka & Fitzmaurice, 2017; Moeller, 2021). The modality of uni-, bi-, and multivariate distributions must be checked before averages or other one-size-fits-all sample coefficients can be expected to represent overall group trends in these distributions (for techniques how to, please see

Haslbeck et al., 2022; Machler, 2021, and Appendix C). The fact that this rarely happens implies that we do not know what one sample's coefficients really represent, whether a group statistic generalized to individuals, or let alone whether they can be logically expected to be invariant across samples.

Further Complications When Deciding About the Replicability and Generalizability

Lack of Criteria to Decide Whether a Finding Is Sufficiently Similar to a Previous One (3.1 in Table 2)

A challenge in determining replicability is the lack of clarity about what exactly it means that two findings are sufficiently similar. How similar do they have to be in order to be considered replicated? Studies on replicability typically use the following information or a combination thereof: confidence intervals or Bayesian credibility intervals (Jacob et al., 2019), power (Simonsohn, 2015; Zwaan et al., 2018), replication Bayes Factor (Zwaan et al., 2018), and effect size (Simonsohn, 2015). Despite these various solutions, many problems remain unsolved. For example, how to calculate power and confidence intervals in the often complicated statistical models used in ILS? Can tests of correlations' equivalence be used to compare within-person correlation coefficients in two individuals? Solutions to such problems occurring with complex data with multilayered sources of variation are only about to be developed (Deffner et al., 2021).

Legitimate Researcher Degrees of Freedom in Decisions About Analyses and Inferences (3.2 in Table 2)

Not every failure to confirm a meaningful finding in a new sample is due to methodological problems or a theory's boundary conditions. Sometimes researchers make slightly different decisions on how to collect and analyze data, leading them to different but similarly valid paths of insight than another study examining the same (Gelman & Loken, 2014; Manapat et al., 2022). For instance, the multi-analysts study by Bastiaansen et al.

(2020) asked twelve researchers with experience in the analyses of intensive longitudinal data to analyze the same intensive longitudinal dataset of one person to determine which of the several assessed psychopathological symptoms should be targeted in a treatment. Despite of working on the same data with the same research question, the analytical approaches and conclusions of the multiple analysts differed in various stages of the analyses, including pre-processing, model selection, and treatment recommendations. The authors concluded that the selection of treatment targets in intensive longitudinal studies on psychopathology currently depends on the use of researcher degrees of freedom. The impacts of such legitimate researcher degrees of freedom on the generalizability of findings across research groups needs to be further studied with multi-analysts studies (Aczel et al., 2021; Bastiaansen et al., 2020; Silberzahn et al., 2018) and multiverse analyses (Dragicevic et al., 2019; Steegen et al., 2016). Multi-analyses studies typically analyze the same dataset with the same research question by several independently working researchers to examine (in-)variance across researchers and the decisions they make using legitimate researcher degrees of freedom (e.g., when making decisions about outlier removal, model selection, interpretation of findings, etc.). Thus, multi-analyst studies examine objectivity (in terms of agreement between researchers) regarding data analysis and inferences, providing insights about generalizability across these methodological conditions. Multiverse analysis is „a philosophy of statistical reporting where paper authors report the outcomes of many different statistical analyses in order to show how fragile or robust their findings are" (Dragicevic et al., 2019, p. 2). Multiverse analyses compare the findings between all datasets that result from the different decisions that researchers can reasonably make with their justified researcher degrees of freedom. Where traditional studies report one procedure of data processing as if it was the only possible one (e.g., "we identified and removed outliers with method X, removed noncompliant participants with method Y and chose model Z"), multiverse analyses examine whether such data processing decisions affect the research finding. Thus, multiverse analyses comparing the

invariance of research findings across different data processing procedures provides insights about the generalizability of research findings across different methods that are assumed to do the same. In our framework (Table 1) this is a matter of robustness check (because the same original data is examined with competing analyses to examine invariance across analytic choices).

Confusing a Lack of Replicability With a Lack of Generalizability (3.3. in Table 2)

Failing to observe a previously reported finding in a novel group of individuals, time period, or contextual setting neither necessarily implies non-replicability, nor universal non-generalizability of that previous finding. This is because logically, the finding can be replicable while its generalizability may be limited to a hitherto unknown set of boundary conditions that were present in the original sample but absent in the sample used in the replication attempt (which we call local generalizability).

While boundary conditions are technically an issue of generalizability (limiting the generalizability from one condition to the next), their unknown presence may look like a lack of replicability to the researcher (hidden moderator problem). If a new study fails to confirm a previous finding, this may be due to method issues, the presence of unknown boundary conditions, such as population-, region- or country-specific factors, or researcher degrees of freedom. All of the possibilities described in Table 2 must be considered to understand why exactly an original finding was not observed in a subsequent study.

Unexplained Non-Replicability (4 in Table 2)

Finally, a previous finding may fail to replicate in new samples even after all methodological issues and imaginable boundary conditions were ruled out. No method is perfect, but we should distinguish between flawed practices (below state-of-the-art) and imperfect methodological representations of the truth due to the truth being more complex than our methods can reflect (see Manapat et al., 2022). Even when using state-of-the-art

methods and understanding the here discussed sources of heterogeneity, studies will occasionally lead to spurious findings. For instance, we cannot completely avoid both type I and type II errors.

In this section, we described several conditions that can limit both the replicability and the generalizability of research findings. In the next sections, we discuss an epistemological framework that can guide the steps that researchers can take to determine which of these boundary conditions may account for variation of a finding between studies.

Applying Lakatos's Concept of Sophisticated Methodological Falsificationism to the Question of not just Whether, but Under what Conditions Research is Replicable and Generalizable

Does a failed replication attempt showing variation of a research finding between two samples logically necessarily imply this finding's non-replicability and non-generalizability? We propose that depending on the sources of such variance between samples, research findings can be replicable and even in certain aspects generalizable. Below we propose a framework for how to define and determine replicability and generalizability in the presence of heterogeneity.

Thesis 1: That research findings should be replicated before being trusted has been consensus even before the current replicability debate, but that generalizability should be equally tested before being claimed is not yet established to the same degree (Yarkoni, 2022a). Many authors discuss their findings as if they were universally valid without any systematic research on possible sample-, time-, or context-specificity, seemingly applying the principle that "When a universal property of nature or biology is being explored, generalizability is often assumed" (Goodman et al., 2016, p. 1). However, many psychological phenomena are less homogeneous than laws of nature described in physics or

biology (Bryan et al., 2021; Molenaar, 2014; Müller et al., 2019; Richters, 2021). Generalizability should, therefore, be systematically tested before being claimed (Czibor et al., 2019; Yarkoni, 2022a; 2022b). This could help solving the current practice in which homogeneity / global generalizability is typically assumed but this hypothesis rarely tested and counter-evidence systematically ignored or not even detected, even if it is captured in the collected data, which often enough it is not (see Figure 2).

Thesis 2: Generalizability cannot be logically proven, because we cannot rule out that hitherto unknown boundary conditions may be found later (because the absence of evidence is not evidence of absence; Wright, 1888, p. 59; see the hidden moderator debate, Wingen et al., 2020). Although we need to systematically seek to identify boundary conditions limiting either the scope of our studied theory, or our ability to correctly infer replicability and generalizability, this process of scrutinizing our theories and methods is never complete. That we can never rule out the existence of boundary conditions should not let us despair and give up. It should encourage us to narrow down as much as possible on the scope of validity of our theories, and the trustworthiness of our inferences of replicability and generalizability.

Thesis 3: Before examining all the points 1.1 through 2.5 from Table 2, it is impossible to know whether a finding differed between an original study and a replication attempt due to being truly spurious in either study, or if the finding was replicable as such, but limited to boundary conditions that differed between the original and replication studies without the researchers knowing it.

To disentangle non-replicability from non-generalizability and to understand either one's causes better, we need to examine step by step which of the issues 1.1-2.5 in Table 2 affected the findings. We propose to apply Lakatos' concept of *sophisticated methodological falsificationism* to the question not just whether, but under which circumstances (boundary conditions) original findings replicate and generalize. The conditions of replicability and generalizability should be studied with systematic research programs in terms of Lakatos

(1978), including systematic explorations of all the sources of heterogeneity mentioned in Table 2 (for practical recommendations of how to explore boundary conditions, see e.g., Busse et al., 2017; Golino et al., 2022; Yarkoni. 2022a, and Appendix C). While Lakatos' systematic research programs originally referred to studies of the boundary conditions of a *theory's* scope of validity, we propose to also apply Lakatos' search for boundary conditions to our *methods' and research practices'* scopes of validity (because measurement theory and philosophy of science are theories, too). This implies for instance that the generalizability of a causal mechanism or a measurement model across individuals, time points, or other context factors should be explored and established empirically by comparing their invariance across these factors (see e.g., Yarkoni. 2022a; Flake et al. 2022).

Thesis 4: To facilitate the narrowing down on boundary conditions in systematic research programs in terms of Lakatos (1978), possible and plausible boundary conditions should be addressed and reported more transparently and systematically in research publications (for solutions, see e.g., Bryan et al., 2021; Busse et al., 2017; Pearl & Bareinboim, 2014; Rohrer et al., 2021; Simons et al., 2017 and Appendix C).

Thesis 5: Generalizability is no binary concept fully falsified if one boundary condition is identified. Instead, it can be insightful to distinguish between *universal/global versus local generalizability* (Brandtstädter, 1985; Czibor et al., 2019). *Universal/global generalizability* is assumed (until contrarian evidence turns up) if a finding is invariant across all samples and subgroups without any obvious boundary conditions (so far) limiting its existence to subgroups of individuals, contexts, or time points. *Local generalizability* is assumed (until further notice) if an effect found in one context is observed under certain boundary conditions but not others. Local generalizability would be, for instance, confirming an original person-level (inter-individual) effect in a different population of individuals (e.g., a different country) but only in individuals with a certain demographic characteristic (e.g., university education). Local generalizability requires a specification such as "This finding is observed (generalized)

across the following conditions A, B, and C (e.g., the theory predicts empirical observations in Germany, the US and Mexico), but only if the conditions D, and E are met (e.g., only if the individuals are at least ten years old and not affected by mental or physical health issues)."

Thesis 6: Most studies empirically testing generalizability follow a hypothetico-deductive logic. Starting with the hypotheses that their findings are generalizable, and then falsifying this hypothesis by contrasting it with empirical evidence revealing certain boundary conditions.This is usually guided by theory-derived hypotheses about plausible boundary conditions. An alternative, *inductive* approach is to first examine the specific findings (correlation or regression coefficients) in individual units of analysis (e.g., individual persons). Then using certain inductive (often machine learning) procedures (e.g., GIMME method: Beltz et al., 2016; superlearning: Luedtke & van der Laan, 2016; Montoya et al., 2021) to identify recurrent patterns that generalize across these units (here: individuals), and represent nomothetic laws that generalize across individuals. In this inductive approach, generalizability *is established empirically before being claimed*. This bottom-up, data-driven approach (e.g., Busse et al. 2017; Yarkoni, 2022a) has been developed as a solution to determining generalizability in the face of heterogeneity and lacking ergodicity (see also Golino et al., 2022; Nesselroade & Molenaar, 2010). It avoids the problem of one-size-fits-all coefficients overlooking heterogeneity in mixture distributions addressed in 2.4 in Table 2 (see also Moeller, 2021). Various methodologists consider the practice of merely assuming but not testing nomothetic, generalizable laws to be a failed epistemological paradigm (a degenerating research program, in Lakatos' terms; see Bryan et al., 2021; Moeller, 2021; Molenaar, 2004; Richters, 2021). Considering that, it seems worthwhile to consider such inductive idiographic-to-nomothetic approaches as useful complements to theory-driven tests of hypothesized boundary conditions. Inductive and deductive, idiographic and nomothetic, approaches can be integrated (e.g., Beck & Jackson, 2020; Beltz et al., 2016; Moeller et al., 2022a), for instance by combining theory-guided approaches with cluster/subgroup analyses,

or by using induction in the phase of hypothesis-generation and subsequent deductive approaches in hypothesis-testing.

Thesis 7: Arguably, some of the problems currently being debated might have been solved or avoided altogether if certain insights from past epistemological debates had not fallen into oblivion. The recently proposed integrations of idiographic and nomothetic approaches (Beltz et al., 2016), of inductive and hypothetico-deductive approaches (Moeller et al., in prep.), and our reminder of Lakatos' (1978) concept of sophisticated methodological falsificationism all build upon debates that took place many decades ago. Including the debates among Popper, Lakatos and Kuhn (for a summary, see Andersson, 2019), the early debates about the advantages and limitations of the hypothetico-deductive research logic and solutions for iterative oscillations between inductive and deductive approaches (Adorno et al., 1976; Glaser & Strauss, 1967) and the idiographic research philosophy (Windelband, 1894/1998; Stern, 1911). Maybe our research community could have prevented some of the masquerading of exploratory findings as confirmatory ones (i.e., the p-hacking and HARKing), had it acknowledged earlier the potential usefulness of inductive approaches that were discussed in past epistemological debates and method development (e.g., Adorno et al., 1976; Glaser & Strauss, 1967; Moeller et al., 2022a; Tracy, 2012).

The debates about the various crises in psychology (theory; measurement; construct validity; normativity; inferential; replicability; and generalizability crisis, for references see Table 2 and Appendix B) could benefit much from a better understanding of previous epistemological debates. As Lakatos (1971, p. 91) put it, "Philosophy of science without history of science is empty; history of science without philosophy of science is blind". Therefore, we would like to remind of epistemological works that have early on pointed out that boundary conditions need to and can be studied systematically, to gain knowledge about the scope to which a theory does or does not apply (i.e., the boundaries within its statements generalize). Let's have and publish more epistemological arguments.

Relevance of This Article for the Research on Replicability and Generalizability in Personality and Social Psychology

The challenges to replicability and generalizability described in the points 2.1 through 3.5 in Table 2 are particularly relevant in the work with intensive longitudinal data (Moeller et al., in prep.). However, they are also more broadly relevant for much of the research in Social and Personality Psychology:

Intensive longitudinal data are becoming more and more frequent and more important in many research fields (Hamaker & Wichers, 2017), including Social and Personality Psychology (Hofmans et al., 2019), by which the latter research fields inherit the challenges to replicability and generalizability described in the right panel of Figure 1. For instance, the person-situation debate (both the original and its recent reiteration; Mischel, 1979; Kenrick & Funder, 1988; Fleeson & Noftle, 2008; Roberts & Caspi, 2001) has led to calls for more studies on situation-level variability and within-person patterns (Fleeson, 2004), which has brought all the problems of time-specificity, context-specificity and heterogeneity into personality research (Baumert et al., 2017; Leikas et al., 2012).

In part due to these calls for studies on within-person variability, topics such as lacking ergodicity and between-person heterogeneity in regard to within-person patterns have become increasingly relevant in research on Personality and Social Psychology (Dotterer et al., 2020; Molenaar, 2014; Moeller, 2021; Fleeson & Noftle, 2008; Itzchakov et al., 2022; Riccio et al., 2019; Weinstein et al., 2022).

In addition to intensive longitudinal data becoming more salient and important in the Social and Personality Psychology research, the challenges to replicability and generalizability depicted in the the points 2.1 through 2.5 in Table 2 are also highly relevant for these research fields. For instance, context-specificity is studied in research on people's everyday thoughts (Baumeister et al., 2020) and everyday trust (Weiss et al., 2021). Time-

specificity is studied in research on personality development and studies addressing person-situation interactions (e.g., Fleeson, 2007), and heterogeneity invalidating averages and other measures of central trends and idiographic approaches are discussed in research on personality and inter-individual differences (Conner et al., 2009; Moeller, 2021; Molenaar, 2004).

Different current debates in psychology calling for systematic studies of boundary conditions to establish generalizability

To implement the proposed Lakatosian (1978) systematic research program examining the conditions of replicability and generalizability, we remind of Lakatos' concepts, especially *progressive programs* (altered assumptions helping to predict *new* phenomena), and *degenerating programs* (auxiliary hypotheses failing to improve predictions of new phenomena; see also Meehl, 1990). A research program examining these auxiliary hypotheses[5] step by step is progressive as long as it reveals new insights about sources of variation or method artifacts influencing the invariance of the finding across samples (or analysts, or analyses, etc.). A (for the moment) degenerating research program studying replicability and generalizability is one failing to cumulate knowledge about the exact conditions under which a finding can be replicated and generalized, such as much psychological research before the replicability debate had been (Ioannidis, 2005; 2012; Singh, 2022).

To better understand the sources of heterogeneity limiting the replicability and generalizability of research findings, Bryan et al. (2021, p. 2) call for the following measures

[5] Please note that by referring to auxiliary hypotheses, our purpose is not to help researchers immunize their theories against falsification by explaining away cases of non-replicability. Instead,we aim to shed light on and explicitly discuss boundary conditions of each study. Testing theories by throwing all the counter-arguments and counter-evidence at them matters to us as to anyone else. A discussion of a theory's boundary conditions is nevertheless both theoretically and practically relevant.

to promote what the authors call the *heterogeneity revolution*: "increased attentiveness, in the hypothesis generation phase, to the likely sources of heterogeneity in treatment effects [or any other sample coefficient]; (2) efforts to measure characteristics of samples and research contexts that might contribute to such heterogeneity; (3) the use of new, conservative statistical techniques to identify sources of heterogeneity that might not have been predicted in advance; and, ultimately, (4) large-scale investment in shared infrastructure to reduce the currently prohibitive cost to individual researchers of collecting data—especially field data—in high-quality generalizable samples."

When pursuing measures as just outlined, researchers need to dodge the risk of forking, which is the problem that in a maze of possible explanations, all branches need to be examined. Otherwise one explored branch may appear plausible while three unexplored paths may have been more insightful (Gelman & Loken, 2014). Multiverse analyses and many-analysts studies, both examining heterogeneity in data analyses due to researcher degrees of freedom, can help overcome forking (Aczel et al., 2021; Bastiaansen et al., 2020; Dragicevic et al., 2019; Silberzahn et al., 2018; Steegen et al., 2016; Weermeijer et al., 2022).

Additional practical solutions for improving the generalizability of research findings have been proposed in the debate about the *generalizability crisis* (e.g., Visser et al., 2022; Yarkoni, 2022b; see also Appendix C). We aspired to link these efforts to the discussion about sources of heterogeneity in articles about the *heterogeneity revolution* (Bryan et al., 2021) and limited generalizability of group trends (e.g., averages) to individuals and subgroups (e.g., Lundh, 2022; Moeller, 2021; Molenaar, 2004). Figure 2 explains how and with what arguments and methodological innovations current debates in Psychology call for a systematic exploration of boundary conditions, and how that relates to the traditional nomothetic and hypothetico-deductive approach. Appendix C summarizes practical research approaches that have been proposed to systematically study boundary conditions and to understand and mitigate reasons for non-generalizability.

Figure 2: How current debates in Psychology innovate traditional epistemological approaches by calling for explorations of boundary conditions

The traditional approaches:

Nomothetic approaches	**Hypothetico-deductive research logic in combination with falsificationism**
One law of behavior expected to apply to all (Wundt/Windelband) or most (Galton) individuals. Typically examined with group-based statistics, such as averages, average treatment effects, etc. Central assumptions: Effects do not vary across individuals (Windelbandt), or if they do, that variance is noise normally distributed around the true effect (Galton), which is typically inferred from an average.	Starts with a theory, derives a hypotheses, tries to falsifiy the null hypothesis. Very cautious about of auxiliary hypotheses stating boundary conditions, concern is that they may be used to immunize the core of a theory against falsification („Lakatos' challenge'). Typically tests universal statements (e.g., „There is no gender difference in outcome X"). A moderator (e.g., gender affecting an outcome) is usually described in the alternative hypothesis.

The hypothetico-deductive falsification works well together with the nomothetic approach, because the nomothetic assumptions of universal generalizability can logically be falsified and are therefore insightful to test.

Recent debates questioning aspects of the traditional approaches:

Debate about Generalizability crisis	**Debate about heterogeneity revolution**	**Debates about lacking ergodicity and within-person (person-oriented) analyses**	**Debate about limitation of nomothetic approaches and novel methodological integrations of nomothetic with idiographic approaches**	**Role models of more pragmatic integrations of inductive and deductive research steps in other disciplines and psychological subfields**
Discusses the problem that research findings often vary between many different conditions, including variance across: *1. Different methods* 1.1 operationalizations and measures that are theoretically expected to capture the same, 1.2 Different analyses that are theoretically expected to be equivalent *2. Sample characteristics* *2.1* Individuals, samples, populations (the latter point is also addressed in the debate about WEIRD samples: Henrich et al., 2010). 2.2 Time points 2.3 Contexts Led to proposals for innovative methods and research strategies that study and reveal sources of invariance limiting generalizability across the above-mentioned conditions.	Shows that heterogeneity is prevalent and that it limits the generalizability across many boundary conditions, such as those addressed in debates about the generalizability crisis. Calls for systematic *explorations of* (i.e., inductive research on) boundary conditions. Criticizes that the hypothetico-deductive research paradigm had lead to researchers not refusing to embrace hypothesies about boundary conditions (argument: immunization against falsification of theory core by adding auxiliary hypotheses), not assessing sources of heterogeneity, not analyzing variance across boundary conditions, and downplaying Raises the concern that such heterogeneity across boundary conditions is in logical conflict with the assumption of homogeneity underlying nomothetic approaches, which by definition expects either invariance between individuals, or expects variance in form of unimodal distributions in which random noise leads to ignorable variation of empirical effects clustering around the true effect.	Points out why and how often group-based statistics fail to describe the individuals in that group (e.g., why an average may fail to describe anyone in the sample in a multi-modal / mixture distribution). Implies that the methods commonly used to test nomothetic hypotheses may fail to generalize to many or even every individual in that sample. Therefore demands that theories supposed to apply to real individuals should be studied with methods that generalize to individuals. Explains that lacking ergodicity implies a lack of generalizability of conclusions obtained with group-based analyses to heterogeneous subgroups or individuals.	Has demonstrated that a top-down (deductive) assumption of nomothetic laws can fail to describe every single individual in a sample, whereas a bottom-up (inductive) identification of inter-individually generalizable findings can do a better job at describing both idiographic and nomothetic findings meaningfully. Has shown that idiographic and nomothetic approaches can be integrated: For instance, powerful machine learning tools can describe both individual patterns (idiographic) and reveal which of these patterns generalize across individuals (nomothetic).	e.g. Data science using inductive data mining and exploratory techniques to generate hypotheses, iterating between hypothesis generation in training datasets and hypothesis testing in test datasets. See also the sophisticated ways in which qualitatively working social scientists try to integrate hypothesis generation and hypothesis testing while allowing for exploration and testing of boundary conditions reveaing heterogeneity across individuals and other conditions.

Conclusions: The recent debates mentioned above imply that the nomothetic assumption of homogeneity is often empirically wrong, but that many studies are unable to detect that. Contributing to this systematic blindness towards existing boundary conditions is the way how many researchers implement the hypothetico-deductive epistemology by allowing researchers to simply assume and claim generalizability a priori, without providing any tests thereof. A consequence is that many researchers see no reason to theorize, hypothesize about or study boundary conditions. Data collections tend to capture limited boundary conditions and are often systematically biased towards some of them (see the discussion about WEIRD samples). Such data are analyzed with methods that are unable to detect heterogeneity and logically assume homogeneity, and the resuls are then typically interpreted as generalizable across individuals, populations (nomothetic approach..) and other boundary conditions. Together, this creates systematic blindness towards heterogeneity and boundary conditions. It implies that many studies following the prevailing version of nomothetic, hypothetico-deductive research often overlook the evidence of heterogeneity that would falsify a nomothetic assumption, and that the resulting misinterpretations of non-generalizable findings as generalizable ones have practical negative consequences for the trustworthiness, generalizability and applicability of psychological findings.

The above-mentioned debates also show how some careful inductive approaches (integrated with hypothetico-deductive approaches searching for nomothetic, i.e., inter-individually generalizable findings) can avoid such misunderstandings. In sum, the debates lead to the conclusion that nomothetic statements are possible, but only trustworthy if they are tested and not just stated. These tests have to include a systematic exploration of boundary conditions. Generalizability is what we call it if after testing all boundary conditions coming to our mind we find no invariance of our findings across these conditions.

As Figure 2 summarizes, several current debates culminate in calls for a systematic exploration of boundary conditions, including debates about the generalizability crisis, heterogeneity revolution, lacking ergodicity and its implications of limited generalizability of between-person results to individuals. These debates together imply that generalizability cannot be properly understood and achieved if we use the hypothetico-deductive research approach as an argument to prevent additional inductive, data-driven explorations of the limitations of generalizability. Whether the hypothetico-deductive research approach logically forbids such explorations, or if only the current practice and interpretation of it prevents such systematic explorations, is a question for further epistemological debates. We believe that using induction for hypothesis-generation and deduction for hypothesis-testing enables us to integrate both approaches without logical inconsistencies and without having to claim mutually exclusive superiority of one approach over the other (see also Moeller et al., 2022a).

Likewise, generalizability cannot be properly understood and achieved if nomothetic (i.e., globally generalizable) statements continue to be claimed without being tested. The search for nomothetic laws has brought many interesting insights. Who does not love the Stroop effect (Stroop, 1935), the McGurk effect (McGurk & MacDonald, 1976), or the video of the gorilla walking in plain sight through a ball game remaining invisible for the onlooker (Simons & Chabris, 1999)? Arguably, the discovery of these findings could be attributed to the nomothetic quest for general effects, although moderators have also since been discussed for each of the effects mentioned here. Despite the important insights gained using a nomothetic approach, some research questions, theories, and practical problems require psychologists to look beyond one-size-fits-all trends. The quest for nomothetic psychological laws has prevented researchers from theorizing and hypothesizing about boundary conditions. For instance, a researcher may feel discouraged from hypothesizing about boundary

conditions in the theory section out of concern that this may decrease the parsimony of their theoretical model. Similarly, a researcher may wish to avoid theoretical speculations about possible boundary conditions in a discussion section, arguing that evidence-based discussion implies that only factors that were examined in the study should be mentioned in the discussion. In addition to the limited reasoning about boundary conditions, samples and data collections typically capture only very limited samples (see e.g., Henrich et al., 2010) and many researchers have made the experience that funding and publications of studies attempting to replicate a previous finding can be met with much reservation (e.g., Bryan et al., 2021). For instance, some of the authors of this article were told that the mere fact that a finding has not been examined in a given country does not justify doing so, or that assuming that a presumed nomothetic hypothesis should be tested in different countries was not parsimonious and a waste of resources. Oftentimes, such debates essentially demand for a good reason to justify the assumption of heterogeneity, and with the dominance of the nomothetic mindset and the discouraged discussion and empirical study of heterogeneity, such good reasons are hard to come by, and harder to get acknowledged.

Even if data collections capture sources of heterogeneity, these boundary conditions are often overlooked in the analyses. Since the predominant analytic methods were all developed for a nomothetic paradigm and therefore look for one-size-fits-all group trends, they tend to systematically overlook heterogeneity (see debates about lacking ergodicity and person-oriented methods). If heterogeneity is accidentally found as an unexpected finding, this may be prevented from being published based on the argument that hypothesizing about unexpected findings is inductive and therefore in (presumable) conflict with the hypothetico-deductive approach. Whether these are logically necessary problems of the nomothetic and hypothetico-deductive research approach, or if the accumulation of these issues represents just bad implementations of these epistemological theories, is a question for further debate.

Integrations of nomothetic and idiographic, hypothetico-deductive and idiographic approaches are possible (see Figure 3).

Figure 3: Contradictions and possible integrations of nomothetic and idiographic, hypothetico-deductive and idiographic approaches

Nomothetic approaches (thesis)	**Idiographic approaches (antithesis)**
Assume invariance across individuals in a group, or consider variance to be random noise, whereas the group trend (e.g., average is considered an indicator for generalizable group trends / nomothetic laws. Study assumed generalizability with methods that look for one-size-fits-all trends (e.g., describe one average treatment effect, one regression coefficient for an entire sample, etc). Aim for descriptions of entire populations and fundamental rules of behavior.	Argue that the methods used to study nomothetic laws, i.e., group trends, can fail to describe some or even all individuals in a sample. That is because they only look for group trends, tend to be blind to heterogeneity and lacking ergodicity, and stop the reasoning process after identifying (or failing to identify) presumable group trends (see the debates about person-oriented methods and heterogeneity revolution). Many methods (e.g., averages) assume homogeneity (normal distribution, unimodality) but that assumption is almost never tested. Point out that understanding and treating individuals is relevant and increasingly so (personalized learning, personalized medicine, etc). Idiographic approaches make that possible, nomothetic ones often do not. Argue that empirical evidence often falsifies nomothetic assumptions if they are tested, which is why just assuming and never testing them is misleading and not insightful.
Integration of idiographic and nomothetic approaches (synthesis)	
Don not stop at stating generalizability across individuals (or other conditions) a priori. Test it and make sure others can understand and test it, too. Establish nomothetic findings bottom-up by empirically identifying patterns that generalize across individuals or other conditions. Use within-person methods to test within-person hypotheses, to avoid overlooking heterogeneity that would invalidate the nomothetic assumption. Test assumptions of methods that require homogeneity. For instance, do not just assume that an average represents the central tendency of a normal distribution, test these assumptions before interpreting averages as indicators of nomothetic group trends. Integrate idiographic and nomothetic approaches, and consider using the power of machine learning to handle the vastness of information about individual models and yet be able to identify generalizable nomothetic patterns that are invariant across many individuals.	

Hypothetico-deductive approaches (thesis)	**Inductive, exploratory approaches (antithesis)**
The order should be build a theory, deduce an a priori hypothesis, test the hypothesis by comparing it with empirical data, trying to find all the counter-arguments and counter-evidence that you (and the community) can come up with. Is cautious towards exploration of boundary conditions. Arguments: auxiliary hypotheses can be used to immunize a theory against falsification. Subgroups identified in exploration may be spurious (i.e., sample-specific, non-generalizable, theoretically or practically irrelevant). Logical certainty of a conclusion about a theory's veracity can only be obtained via falsification, says Popper. But he also says that, even when using falsification, a researcher who really wants to always finds a way around to defend whatever theory they want.	Data-driven. Theories and hypotheses are formed a posteriori, after looking at data and identifying patterns in these data. Have been proposed to increase our knowledge about boundary conditions (e.g., Busse et al., 2017; Yarkoni, 2022a, 2022b). Make it possible to learn from and follow up on unexpected findings (make the most of discovery). Used by many disciplines to gain valuable insights. Data scientists and AI applications use it and outperform any theory-driven prediction. Children use it to learn language. Who are we as social scientists to say no insight can come from induction?
Integration of hypothetico-deductive and inductive approaches (synthesis)	
Use inductive exploration for hypothesis generation and hypothetico-deductive testing for hypothesis testing. See for instance data scientists' iteration between finding patterns in training datasets and testing the generalizability of these findings in test datasets. Theorize and hypothesize about boundary conditions and then test them, but also leave space and allow for surprising, unexpected finding to teach you valuable lessons about the limitations of your theoretical expectations. Whether you can use deduction or need exploration depends a bit on the available prior knowledge about boundary conditions. The less is known, the more exploration may be useful (see Busse et al., Table 3 for that). Induction does not rule out prior theory testing. For instance, there can be 5 theory-testing studies, resulting in 5 datasets. Why not use Directed Acyclical Graphs to develop hypotheses about the causal mechanisms that would be most in line with the covariance previously observed in these studies? This can be integrated in the theory and further tested, until such tests teach us nothing new and valuable (i.e., as long as the research program is progressive, in Lakatos' terms). Ask where good theories come from and if any hypothesis is ever really strictly a priori (before knowing any evidence). Even an intuition reflects empirical experience with the world. Why not be more honest about that and acknowledge that if we are already using empirical knowledge when creating hypotheses, we can be more systematic about that and use data in the theory creation process in ore sophisticated ways (e.g., pattern finding with machine learning), instead of just using our unsophisticated intuition.	

For now, we can observe that empirical evidence and current debates emphasize the prevalence and practical relevance of heterogeneity. Additionally, many of the practical solutions that have been recently suggested transcend the nomothetic and hypothetico-deductive approaches by integrating data-driven explorations and person-specific models into quests searching for ways how to test theories and how to establish knowledge about generalizable findings and the limitations thereof. In any case, we should avoid the naïve falsification (to use Lakatos' terms) consisting of declaring a research finding to be non-replicable and non-trustworthy simply because it varies between two samples (see see Figure 1). Without systematically studying the boundary conditions mapped out in Table 2, it is logically impossible to distinguish non-replicability in terms of spurious findings from limited or lacking generalizability due to specific boundary conditions. Practical steps for the systematic study of boundary conditions are summarized in Appendix C.

References

Aczel, B., Szaszi, B., Nilsonne, G., Van den Akker, O., Albers, C., van Assen, M. A., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., Van Dongen, N. N., Donkin, C., Van Doorn, J. B., ... Wagenmakers, E. J. (2021). Science Forum: Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife*, *10*, e72185. https://doi.org/10.7554/eLife.72185

Adorno, T. W., Albert, H., Dahrendorf, R., Habermas, J., Pilot, H., & Popper, K. R. (Eds.) (1976). *The positivist dispute in German sociology*. Heinemann.

Andersson, G. (2019). Karl Popper und seine Kritiker: Kuhn, Feyerabend und Lakatos [Karl Popper and his critics: Kuhn, Feyerabend and Lakatos]. In G. Franco (ed.), *Handbuch Karl Popper* [Handbook Karl Popper] (pp. 717–731). Springer. https://doi.org/10.1007/978-3-658-16239-9_52

Association for Computing Machinery. (2020, August 24). *Artifact Review and Badging - Current Artifact Review and Badging Version 1.1.* ACM. https://www.acm.org/publications/policies/artifact-review-and-badging-current

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior research methods*, *43*(3), 666-678. https://doi.org/10.3758/s13428-011-0089-5

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S.-M., de Jonge, P., Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E. L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravecz, Z., Riese, H., Rubel, J., … Bringmann, L. F. (2020). Time to get personal? The impact of researchers' choices on the selection of treatment targets using the experience sampling

methodology. *Journal of Psychosomatic Research*, *137*, 110211. https://doi.org/10.1016/j.jpsychores.2020.110211

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences, USA, 113(27), 7345–7352.

Baumeister, R. F., Hofmann, W., Summerville, A., Reiss, P. T., & Vohs, K. D. (2020). Everyday Thoughts in Time: Experience Sampling Studies of Mental Time Travel. *Personality and Social Psychology Bulletin, 46*(12), 1631–1648. https://doi.org/10.1177/0146167220908411

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., Costantini, G., Denissen, J. J. A., Fleeson, W., Grafton, B., Jayawickreme, E., Kurzius, E., Macleod, C., Miller, L. C., Read, S. J., Roberts, B., Robinso, M. D., Wood, D., & Wrzus, C. (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality, 31*(5), 503–528. https://doi.org/10.1002/per.2115

Beck, E. D., & Jackson, J. J. (2020). Consistency and change in idiographic personality: A longitudinal ESM network study. *Journal of Personality and Social Psychology*, *118*(5), 1080–1100. https://doi.org/10.1037/pspp0000249

Beltz, A. M., Wright, A. G., Sprague, B. N., & Molenaar, P. C. (2016). Bridging the nomothetic and idiographic approaches to the analysis of clinical data. Assessment, 23(4), 447-458.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407–425. https://doi.org/10.1037/a0021524

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117*(2), 187–215. https://doi.org/10.1037/0033-2909.117.2.187

Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, *148*(4), 601–618. https://dx.doi.org/10.1037/xge0000558

Bollen, J. T., Cacioppo, R., Kaplan, J., Krosnick, J. L., & Olds, J. L. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences.* www.nsf.gov%2Fsbe%2FAC_Materials%2FSBE_Robust_and_Reliable_Research_Report.pdf&usg=AOvVaw2nUYV1UuFM-Mc1zraiXfJ7

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2015). Harking's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69*(3), 709–750. https://doi.org/10.1111/peps.12111

Brandtstädter, J. (1985). Individual development in social action contexts: Problems of explanation. In J. R. Nesselroade & A. von Eye (Eds.), *Individual development and social change. Explanatory analysis* (pp. 243-264). Academic Press.

Bringmann, L. F., Albers, C., Bockting, C., Borsboom, D., Ceulemans, E., Cramer, A., Epskamp, S., Eronen, M. I., Hamaker, E., Kuppens, P., Lutz, W., McNally, R. J., Molenaar, P., Tio, P., Voelkle, M. C., & Wichers, M. (2022). Psychopathological networks: Theory, methods and practice. *Behaviour Research and Therapy, 149,* Article 104011. https://doi.org/10.1016/j.brat.2021.104011

Bringmann L.F., Elmer T., Eronen M.I. (in press). Back to basics: the importance of conceptual clarity in psychological science. *Current Directions in Psychological Science.*

Bringmann, L., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology 26*(1), 27–43. https://doi.org/10.1177/0959354315617253

Brose, A., Voelkle, M. C., Lövdén, M., Lindenberger, U., & Schmiedek, F. (2015). Differences in the between-person and within-person structures of affect are a matter of degree. *European Journal of Personality, 29*(1), 55–71. https://doi.org/10.1002/per.1961

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour, 5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Busse, C., Kach, A. P., & Wagner, S. M. (2017). Boundary conditions: What they are, how to explore them, why we need them, and when to consider them. *Organizational Research Methods, 20*(4)*,* 574-609. https://doi.org/10.1177/1094428116641191

Casini, E., Richetin, J., Preti, E., & Bringmann, L. F. (2020). Using the time-varying autoregressive model to study dynamic changes in situation perceptions and emotional reactions. *Journal of Personality, 88(4),* 806-821. https://doi.org/10.1111/jopy.12528

Conner, T. S., Tennen, H., Fleeson, W., & Barrett, L. F. (2009). Experience sampling methods: A modern idiographic approach to personality research. *Social and Personality Psychology Compass, 3*(3)*, 292-313*. https://doi.org/10.1111/j.1751-9004.2009.00170.x

Czibor, E., Jimenez-Gomez, D., & List, J.A. (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal*, *86*(2), 371–432. https://doi.org/10.1002/soej.12392

Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A causal framework for cross-cultural generalizability. *Advances in Methods and Practices in Psychological Science, 5*(3)*,* 1–18. https://doi.org/10.1177/25152459221106366

Derrible, S., & Ahmad, N. (2015). Network-based and binless frequency analyses. *PloS One*, *10*(11), e0142108. https://doi.org/10.1371/journal.pone.0142108

Dietrich, J., Schmiedek, F., & Moeller, J. (2022). Academic motivation and emotions are experienced in learning situations, so let's study them. Introduction to the special issue. *Learning and Instruction*, 101623. https://doi.org/10.1016/j.learninstruc.2022.101623

Dotterer, H. L., Beltz, A. M., Foster, K. T., Simms, L. J., & Wright, A. G. (2020). Personalized models of personality disorders: Using a temporal network method to understand symptomatology and daily functioning in a clinical sample. *Psychological Medicine, 50*(14), 2397–2405. https://doi.org/10.1017/S0033291719002563

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019, May). Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300295

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science, 16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Fanelli, D., Costas, R., & Larivière, V. (2015). Misconduct policies,academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLoS One, 10*, e0127556. https://doi.org/10.1371/journal.pone.0127556

Feest, U. (2019). Why replication is overrated. *Philosophy of Science*, *86*(5), 895–905. https://doi.org/10.1086/705451

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science, 12*(1), 46–61. https://doi.org/10.1177/1745691616654458

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393 Flake, J. K., Luong, R., & Shaw, M. (2022). Addressing a crisis of generalizability with large-scale construct validation. *Behavioral and Brain Sciences, 45*, e14. https://doi.org/10.1017/S0140525X21000376

Fleeson, W. (2004). Moving personality beyond the person-situation debate: The challenge and the opportunity of within-person variability. *Current Directions in Psychological Science*, *13*(2), 83–87. https://doi.org/10.1111/j.0963-7214.2004.00280.x

Fleeson, W. (2007). Using experience sampling and multilevel modeling to study person-situation interactionist approaches to positive psychology. In A. D. Ong & M. H. M. van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (pp. 501–514). Oxford University Press.

Fleeson, W., & Noftle, E. (2008). The end of the person–situation debate: An emerging synthesis in the answer to the consistency question. *Social and Personality Psychology Compass, 2*(4), 1667–1684. https://doi.org/10.1111/j.1751-9004.2008.00122.x

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Perceived learning environment and students' emotional experiences: A multilevel analysis of mathematics classrooms. *Learning and Instruction, 17*(5), 478-493. https://doi.org/10.1016/B978-012372545-5/50003-4

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–465. https://doi.org/10.1511/2014.111.460

Georgas, J., van de Vijver, F. J. R., & Berry, J. W. (2004). The ecocultural framework, ecosocial indices, and psychological variables in cross-cultural research. *Journal of Cross-Cultural Psychology*, *35*(1), 74–96. https://doi.org/10.1177/0022022103260459

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory. Strategies for Qualitative Research*. The Sociology Press.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*, 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Hamaker, E. L. (2012). Why Researchers Should Think Within-Person: A Pragmatic Rationale. In M. R. Mehl and T. S. Conner (Eds.), *Handbook of Research Methods for Studying Daily Life* (pp. 43–61). New York, NY: Guilford.

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177%2F0963721416666518

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L., Altman,

S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), 180448. https://dx.doi.org/10.1098/rsos.180448

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2021). Modeling psychopathology: From data models to formal theories. *Psychological Methods.* Advance online publication. https://dx.doi.org/10.1037/met0000303

Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2022, May 9). Multimodality and Skewness in Emotion Time Series. *PsyArXiv:* https://doi.org/10.31234/osf.io/qudr6

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: are they important?. *Journal of Social Behavior and Personality*, *5*(4), 41–49.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(7302), 29. https://doi.org/10.1038/466029a

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations.* Sage.

Hofmans, J., De Clercq, B., Kuppens, P., Verbeke, L., & Widiger, T. A. (2019). Testing the structure and process of personality using ambulatory assessment data: An overview of within-person and person-specific techniques. *Psychological Assessment, 31*(4), 432–443. https://doi.org/10.1037/pas0000562

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

Itzchakov, G., Reis, H. T., & Weinstein, N. (2022). How to foster perceived partner responsiveness: High-quality listening is key. *Social and Personality Psychology Compass*, *16*(1), e12648. https://doi.org/10.1111/spc3.12648

Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M.-A. (2019). A framework for learning from null results. *Educational Researcher*, *48*(9), 580–589. https://doi.org/10.3102/0013189X19891955

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist, 43*(1), 23. https://doi.org/10.1037/0003-066X.43.1.23

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kievit, R., Frankenhuis, W. E., Waldorp, L., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology, 4*, 513. https://doi.org/10.3389/fpsyg.2013.00513

Klein, K. J., & Kozlowski, S. W. J. (2000). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. San Francisco, CA: Jossey-Bass.

Lakatos, I. (1971). History of science and its rational reconstructions. In R. C. Buck, & R. S. Cohen (Eds.), *PSA 1970. Boston studies in the philosophy of science* (Vol. 8, pp. 91–136). Dordrecht. http://doi.org/10.1007/978-94-010-3142-4_7

Lakatos, I. (1978). *The methodology of scientific research programmes.* Cambridge University Press. https://doi.org/10.1017/CBO9780511621123

Larson, R., & Csikszentmihalyi, M. (2014). The experience sampling method. In M. Csikszentmihalyi (Ed.), *Flow and the foundations of positive psychology* (pp. 21–34). Springer. http://doi.org/10.1007/978-94-017-9088-8_2

Lathia, N., Rachuri, K. K., Mascolo, C., & Rentfrow, P. J. (2013, September). Contextual dissonance: Design bias in sensor-based experience sampling methods. *In Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (pp. 183-192). https://doi.org/10.1145/2493432.2493452

Leikas, S., Lönnqvist, J. E., & Verkasalo, M. (2012). Persons, situations, and behaviors: Consistency and variability of different behaviors in four interpersonal situations. *Journal of Personality and Social Psychology, 103*(6), 1007–1022. https://doi.org/10.1037/a0030385

Luedtke, A. R., & van der Laan, M. J. (2016). Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics, 12*(1), 305–332. https://doi.org/10.1515/ijb-2015-0052 https://doi.org/10.1515/ijb-2015-0052

Lundberg, I., Johnson, R., and Stewart, B. M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review, 86*(3), 532–565. https://doi.org/10.1177/00031224211004187

Lundh, L. G. (2019). The crisis in psychological science, and the need for a person-oriented approach. In J. Valsiner (Ed.), *Social Philosophy of Science for the Social Sciences* (pp. 203–233). Springer. https://doi.org/10.1007/978-3

Lundh, L.-G. (2022) The Central Role of the Concept of Person in Psychological Science, Editorial. *Journal for Person-Oriented Research, 8*(2), 38-42. https://doi.org/10.17505/jpor.2022.24853

Maechler, M. (2021). *diptest: Hartigan's Dip Test Statistic for Unimodality – Corrected*. https://github.com/mmaechler/diptest

Manapat, P. D., Anderson, S. F., & Edwards, M. C. (2022). A revised and expanded taxonomy for understanding heterogeneity in research and reporting practices. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000488

Matejka, J., & Fitzmaurice, G. (2017, May). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)* (pp. 1290–1294). Association for Computing Machinery. https://doi.org/10.1145/3025453.3025912 *For the figures mentioned in our Appendix C, please see the blog version (https://www.autodesk.com/research/publications/same-stats-different-graphs), because the Figure numbers differ compared to the pdf manuscript.*

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature 264*, 746–748. https://doi.org/10.1038/264746a0

McManus, R. M., Young, L., Sweetman, J. (2022). Psychology is a Feature of Persons, Not Averages or Distributions: The Group-to-Person Generalizability Problem in Social Cognition Research. *Pre-print*: https://rmmcmanusblog.files.wordpress.com/2023/01/ampps-22-0036.r1_proof_hi.pdf

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115. https://doi.org/10.1086/288135

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Mischel, W. (1979). On the interface of cognition and personality: Beyond the person–situation debate. *American Psychologist, 34*(9), 740. https://doi.org/10.1037/0003-066X.34.9.740

Moeller, J. (2021). Averting the next credibility crisis in psychological science: Within-person methods for personalized diagnostics and intervention. *Journal for Person-Oriented Research, 7*(2), 53–77. https://doi.org/10.17505/jpor.2021.23795

Moeller, J., Bergmann, C., Stockinger, K., Neubauer, A., Schmiedek, F., Riediger, M., Pekrun, R., Bringmann, L., Bastiaansen, J., and the ManyMoments Consortium (in prep.). ManyMoments - Improving replicability of experience sampling method studies in multi-lab collaborations. *Manuscript in preparation.*

Moeller, J., Dietrich, J., & Baars, J. (revise & resubmit). The Experience Sampling Method in the research on achievement-related emotions and motivation. In: R.C. Lazarides, Hagenauer, H. Järvenoja (Eds.), Motivation and Emotion in Learning and Teaching across Educational Contexts: Theoretical and Methodological Perspectives and Empirical Insights. Routledge.

Moeller, J., Langener, A., Lafit, G., Karhulahti, V.-M., Bastiaansen, J. A., & Bergmann, C. (2022a). The hypository: Registering hypotheses for cumulative science. PsyArXiv: https://doi.org/10.31234/osf.io/5qgj7

Moeller, J., Viljaranta, J., Tolvanen, A., Kracke, B., & Dietrich, J. (2022b). Introducing the DYNAMICS Framework of moment-to-moment development in achievement motivation. *Learning & Instruction, 81,* 101653. https://doi.org/10.1016/j.learninstruc.2022.101653

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement:*

Interdisciplinary Research and Perspectives, 2(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Montoya, L., van der Laan, M., Luedtke, A., Skeem, J., Coyle, J., & Petersen, M. (2021). *The optimal dynamic treatment rule superlearner: Considerations, performance, and application*. arXiv. https://doi.org/10.48550/arXiv.2101.12326

Müller, T., Rumberg, A., & Wagner, V. (2019). An introduction to real possibilities, indeterminism, and free will: three contingencies of the debate. *Synthese, 196*(1), 1–10. https://doi.org/10.1007/s11229-018-1842-4

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Murayama, K., Goetz, T., Malmberg, L-E., Pekrun, R., Tanaka, A., & Martin, A. J. (2017). Within-person analysis in educational psychology: Importance and illustrations. In D. W. Putwain, & K. Smart (Eds.), *British Journal of Educational Psychology Monograph Series II: Psychological aspects of education – Current trends: The role of competence beliefs in teaching and learning* (pp. 71–87). Wiley.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour, 3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nesselroade, J. R., & Molenaar, P. C. M. (2010). Emphasizing intraindividual variability in the study of development over the lifespan. In W. F. Overton (Ed.), *The handbook of life-span development: Cognition, biology, and methods across the lifespan* (1st ed., pp. 30–54). Hoboken, NJ: Wiley.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A.

M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review, 26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9), 1046–1058.https://doi.org/10.1016/j.jclinepi.2015.05.029.

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science, 29*(4), 579–595. https://doi.org/10.1214/14-STS486

Pekrun, R., Murayama, K., Marsh, H. W., Goetz, T., & Frenzel, A. C. (2019). Happy fish in little ponds: Testing a reference group model of achievement and emotion. *Journal of Personality and Social Psychology*, *117*(1), 166. https://doi.org/10.1037/pspp0000230

Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, *11*, Article 76. https://doi.org/10.3389/fninf.2017.00076

Plucker, J. A., & Makel, M. C. (2021) Replication is important for educational psychology: Recent developments and key issues. *Educational Psychologist*, *56*(2), 90–100. https://doi.org/10.1080/00461520.2021.1895796

Popper, K. (1935). “Induktionslogik" und “Hypothesenwahrscheinlichkeit” [“Inductive logic and probability of hypotheses”]. *Erkenntnis*, *5*, 170-172. https://www.jstor.org/stable/20011753

Ram, N., Brose, A., & Molenaar, P. C. M. (2013). Dynamic factor analysis: Modeling person-specific process. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 2, pp. 441–457). Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199934898.013.0021

Reitzle, M. (2013). Introduction: Doubts and insights concerning variable- and person-oriented approaches to human development. *European Journal of Developmental Psychology, 10,* 1-8. https://doi.org/10.1080/17405629.2012.742848

Renkewitz, F., & Heene, M. (2019). The Replication Crisis and Open Science in Psychology. *Zeitschrift für Psychologie, 227*(4)*,* 227, 233-236. https://doi.org/10.1027/2151-2604/a000389

Riccio, M. T., Shrout, P. E., Balcetis, E. (2019). Interpersonal pursuit of intrapersonal health goals: Social cognitive–motivational mechanisms by which social support promotes self-regulatory success. *Social and Personality Psychology Compass*, *13*(10), Article e12495. https://doi.org/10.1111/spc3.12495

Richters, J. E. (2021). Incredible utility: The lost causes and causal debris of psychological science. *Basic and Applied Social Psychology*, *43*(6), 366–405. https://doi.org/10.1080/01973533.2021.1979003

Roberts, B. W., & Caspi, A. (2001). Personality development and the person-situation debate: It's déjà vu all over again. *Psychological Inquiry, 12*(2), 104–109. https://doi.org/10.1207/S15327965PLI1202_04

Rohrer, J. M., Schmukle, S., & McElreath, R. (2021). *The only thing that can stop bad causal inference is good causal inference*. PsyArXiv. https://doi.org/10.31234/osf.io/mz5jx

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, *21*(4), 308-320. https://doi.org/10.1037/gpr0000128

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology*, *5*, 1-9. https://doi.org/10.15626/MP.2019.1645

Schloss, P. D. (2018). Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio, 9*(3), e00525-18. https://doi.org/10.1128/mBio.00525-18

Schmiedek, F., Lövdén, M., von Oertzen, T., & Lindenberger, U. (2020). Within-person structures of cognitive performance differ from between-person structures of cognitive abilities. *PeerJ 8e:9290.* https://doi.org/10.7717/peerj.9290

Schmitz, B. (2006). Advantages of studying processes in educational research. *Learning and Instruction, 16(5)*, 433–449. https://doi.org/10.1016/j.learninstruc.2006.09.004

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research on Personality*, *47*, 609-612. https://doi.org/10.1016/j.jrp.2013.05.009

Shadish, W, Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shalit, U., Johansson, F. D., & Sontag, D. (2017, July). Estimating individual treatment effect: generalization bounds and algorithms. *In International Conference on Machine Learning(pp. 3076-3085)*. PMLR. https://doi.org/10.48550/arXiv.1606.03976

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, 337–356. https://doi.org/10.1177/2515245917747646

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28 (9)*, 1059–1074. https://doi.org/10.1068/p281059

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177%2F1745691617708630

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Simonsohn U., Simmons J.P., Nelson L.D. (2020). Specification curve analysis. *Nature Human Behaviour 4,* 1208–1214. https://doi.org/10.1038/s41562- 020-0912-z

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B. 13*, 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

Singh, S. (2022). *Replication data for: political knowledge, the decision calculus, and proximity voting*. Harvard Dataverse. https://doi.org/10.7910/DVN/QA789B

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature, 575*(7781), 9-10. https://doi.org/10.1038/d41586-019-03350-5

Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., ... & Wilson, J. (2019). Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Advances in Methods and Practices in Psychological Science*, *2*(4), 335-349. https://doi.org/10.1177%2F2515245919869583

Stawski, R. S., Scott, S. B., Zawadzki, M. J., Sliwinski, M. J., Marcusson-Clavertz, D., Kim, J., Lanza, S. T., Green, P. A., Almeida, D. M., & Smyth, J. M.. (2019). Age differences in everyday stressor-related negative affect: A coordinated analysis. *Psychology and Aging, 34*(1), 91-105. https://doi.org/10.1037/pag0000309

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Stern, W. (1911). *Die Differentielle Psychologie in ihren methodischen Grundlagen [The differential psychology in its methodological foundations]*. Barth.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651

Syed, M. (2021). *Reproducibility, diversity, and the crisis of inference in psychology*. PsyArXiv. https://doi.org/10.31234/osf.io/89buj

Tracy, S. J. (2012). The toxic and mythical combination of a deductive writing logic for inductive qualitative research. *Qualitative Communication Research, 1*(1), 109-141. https://doi.org/10.1525/qcr.2012.1.1.109

Uygun Tunç, D., & Tunç, M. N. (2020, May 13). A Falsificationist Treatment of Auxiliary Hypotheses in Social and Behavioral Sciences: Systematic Replications Framework. Preprint: https://doi.org/10.31234/osf.io/pdm7y

Vaidyanathan, U., Vrieze, S. I., & Iacono, W. G. (2015). The power of theory, research design, and transdisciplinary integration in moving psychopathology forward.

Psychological Inquiry, 26(3), 209-230. https://doi.org/10.1080/1047840X.2015.1015367

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., Franchin, L., Frank, M. C., Geraci, A., Hamlin, J. K., Kaldy, Z., Kulke, L., Laverty, C., Lew-Williams, C., Mateu, V., Mayor, J., Moreau, D., Nomikou, I., Schuwerk, T., … Zettersten, M. (2022). Improving the generalizability of infant psychological research: the ManyBabies model. *Behavioral and Brain Sciences, 45*. https://doi.org/10.1017/S0140525X21000455

Voelkle, M. C., Brose, A., Schmiedek, F. & Lindenberger, U. (2014). Towards a unified framework for the study of between-person and within-person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research, 49*(3), 193-213. https://doi.org/10.1080/00273171.2014.889593

Weermeijer, J., Lafit, G., Kiekens, G., Wampers, M., Eisele, G., Kasanova, Z., Vaessen, T., Kuppens, P., & Myin-Germeys, I. (2022). Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions. *Behavioral Research Methods*. https://doi.org/10.3758/s13428-021-01777-1

Weinstein, N., Itzchakov, G., & Legate, N. (2022). The motivational value of listening during intimate and difficult conversations. *Social and Personality Psychology Compass*, *16*(2), e12651. https://doi.org/10.1111/spc3.12651

Weiss, A., Michels, C., Burgmer, P., Mussweiler, T., Ockenfels, A., & Hofmann, W. (2021). Trust in everyday life. *Journal of Personality and Social Psychology, 121*(1), 95–114. https://doi.org/10.1037/pspi0000334

Whitaker, K. (2017). *Publishing a reproducible paper*. Figshare. https://doi.org/10.6084/m9.figshare.5440621.v2

Windelband, W. (1894/1998). “History and Natural Science.” *Theory & Psychology, 8*(1): 5–22.

Wingen, T., Berkessel, J. B., & Englich, B. (2020). No replication, no trust? How low replicability influences trust in psychology. *Social Psychological and Personality Science*, *11*(4), 454–463. https://doi.org/10.1177/1948550619877412

Wirtz, M., & Nachtigall, C. (1998). *Deskriptive Statistik. Statistische Methoden für Psychologen [Descriptive statistics. Statistical methods for psychologists]* (2nd ed.). Juventa.

Wright, W., (1888). The empire of the hittites. *Journal of the Transactions of The Victoria Institute, or Philosophical Society of Great Britain, 21*, 55–73.

Yarkoni, T. (2022a). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, 1–37. https://doi.org/10.1017/S0140525X20001685

Yarkoni, T. (2022b). Replies to commentaries on the generalizability crisis. *Behavioral and Brain Sciences, 45.* https://doi.org/10.1017/S0140525X21001758

Zwaan, R., Etz, A., Lucas, R., & Donnellan, M. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, E120. https://doi.org/10.1017/S0140525X17001972

Appendix A

For definitions of replicability and generalizability, please see pages 6-8 in the manuscript. In addition, we would like to clarify how we distinguish the following terms from replicability and generalizability:

The term *reproducibility is used in this article to refer to situations where an independent (new) team of researchers* arrives at the same results and conclusions when they use the same statistical procedures *on the original dataset* (e.g., Schloss, 2018; Whitaker, 2017; see e.g., Hardwicke et al., 2018 for analytical reproducibility rates in cognitive psychology)[6]. We thus limit the term reproducibility to the work within the same original dataset, and the term replicability to the work across different datasets. In other words, reproducibility is a stress-test of the research process on the original dataset, while replicability represents a cross-validation of the findings in a new dataset. Some authors use the term *inferential reproducibility* for the similar concept of drawing similar conclusions

[6] The definitions of reproducibility and replicability by Plesser (2018 and the Association for Computing Machinery (2016) differ from those we adopt from Schloss (2018) and Whitaker (2017). What Plesser (2018) and the Association for Computing Machinery (2016) defined as reproducibility matches the definition of replicability by Schloss (2018) and (Whitaker (2017), and what Plesser (2018) and the Association for Computing Machinery (2016) defined as replicability matches the definition of reproducibility by Schloss (2018) and Whitaker (2017). Note that other researchers have used the term reproducibility to refer to attempts to test the invariance of findings across different datasets (e.g., Goodman et al., 2016), but since this leads to overlaps between definitions of reproducibility and replicability and possible confusion following thereof, we limit the term reproducibility to the work with the original dataset. Please note that Goodman's (2016) distinction between methods replicability, reproducibility, results reproducibility and inferential reproducibility is interesting but somewhat orthogonal to the distinction between reproducibility and replicability applied in this article. Some authors use the terms results reproducibility and replicability interchangeably (Goodman et al., 2016)

from a study's independent replication or reanalysis by an independent team of researchers (Goodman et al., 2016).

Robustness is defined here as finding similar results in the same population and context, but with different methods. An effect is robust with regard to a method of assessment or a method of analysis if two equivalent methods of assessment or analysis lead to similar results, so that method artifacts can be ruled out (Schloss, 2018; Whitaker, 2017).[7]

In regard to all of these definitions mentioned above, a challenge is to define what it means to find "similar results" or draw "similar conclusions". To decide how to distinguish sufficiently similar results and conclusions from insufficiently similar ones, precise decision criteria need to be defined and debated. For instance, how similar do two effects have to be, or should two findings concerning an effect be classified as sufficiently similar if both are significant, or insignificant? While the answer to this question may differ depending on schools of thinking (e.g., Bayesian versus frequentist approaches, see Amar, Shamir, Yekutieli, 2017; Jones, Williams, & McNally, 2020; Savalei & Dunn, 2015; Wagenmakers & Grünwald, 2006), our stance furthermore is that such criteria for deciding about suffice of similarity are in part specific to the methods used in a given study, and still need to be defined and operationalized for studies using intensive longitudinal methods.

[7] Other authors have called this a conceptual replication (e.g., Watts, Duncan, & Quan, 2018), but we limit the term replication as coming to similar conclusions in a *new dataset* (see below), while robustness refers to coming at similar conclusions using *new* analysis on the *original dataset*.

Appendix B

Widely Discussed Method-Related Conditions of (Non-)Replicability

This section describes the methodological reasons for limited replicability addressed in the points 1.1-1.8 in Table 2. All of these issues have been widely discussed, which is why we mention them only in our appendix, for readers who are interested in catching up with these debates.

In the years since the replicability crisis heated up (for a comment on its starting date, see Renkewitz & Heene, 2019), the debate in the scientific community has mostly focused on methodological flaws and flawed research practices leading to untrustworthy findings. We briefly summarize these debates here before pointing out additional, less widely known conditions of (non-)replicability in the subsequent section.

Flaws in concepts, theory and hypothesis formulation ("theory crisis"; 1.1. in Table 2)

The debate surrounding the so-called theory crisis addressed the need for better theories, better hypotheses, and better null hypotheses falsifications (Eronen & Bringmann, 2021; Fiedler, 2017; Fried, 2020; Guest & Martin, 2021; Haslbeck et al., 2021; Muthukrishna & Henrich, 2019; Meehl, 1967; 1990; Oberauer & Lewandowsky, 2019; Smaldino, 2019; Vaidyanathan et al., 2015). The theory crisis addresses for instance the problems that hypotheses are often not sufficiently specific. It is often unclear what kind of evidence exactly would falsify or support them. They are not sufficiently formalized and often stay in the form of vague narrative statements and can be reinterpreted after the results are in.

Lacking Conceptual Clarity About Phenomena and Estimands ("construct validity crisis"; point 1.2 in Table 2)

Much of the psychological research is affected by a lack of clear definitions. For instance, oftentimes, different labels are used to refer to the same phenomenon (jingle fallacy)

or the same label is used to refer to different phenomena (jangle fallacy, see Block, 1995). If definitions are given, they are not always embedded in the existing prior literature and theory development.

In addition to unclear theoretical definitions of the exact phenomenon that a study aims to capture, there is often a lack of explanations on how exactly this theoretical definition is linked to the empirical relative (the data and measurement instruments) that is supposed to represent it (e.g., Bringmann et al., in press; Lundberg et al., 2021)

These problems have been summarized as the construct validity crisis (Schimmack, 2019). They lead to the problem that two studies may attempt to study the same construct but in fact examine two different phenomena because of a lack of clear definitions. This would look like a lack of replicability but in fact is a lack of consensual definitions and construct validity of measurement instruments.

Flaws in the design and instruments of data collections ("measurement crisis"; 1.3. in Table 2)

The debate about the so-called measurement crisis (for overviews, see Bringmann & Eronen, 2016; Flake & Fried, 2020) addressed the problem that measured constructs are often not sufficiently clearly defined to allow precise development of measurement instruments. Also, the criteria why certain measures were selected are often intransparent, many frequently used measures lack validity or at least relevant information about certain aspects of validity. Further issues discussed as characteristics of the measurement crisis include the lack of clear

and sound scoring practices (assigning appropriate numeric relatives to measured phenomena) as well as problematic scale transformations.

Flaws in the sampling procedure of recruiting participants, measurement occasions, and situations (1.4. in Table 2)

In the wake of the replicability debate, it has often been pointed out that small samples pose a risk to replicability because they may lead to untrustworthy significance values and even, in some cases, effect sizes (Schönbrodt & Perugini, 2013). A further challenge to replicability and generalizability are unrepresentative samples: If two samples are not representative, and it is unclear in what exact regards they differ from each other, then we cannot expect them to lead to comparable, replicable, results. This problem is aggravated in intensive longitudinal studies, many of which use samples of convenience / cluster sampling that are limited at least on one level of analyses, for instance, the number of individuals or higher-order levels, such as organizations the individuals are clustered within. In cluster samples of convenience, the known sample characteristics (e.g., locations of work groups or individuals) are often confounded with other sample characteristics that may appear less obvious to the researcher (e.g., rural versus urban environment, socioeconomic status, political or cultural differences between regions). This problem is closely related to the heterogeneity posed by context and person differences that we discuss in points 2.1 and 2.3 in Table 2.

Flaws in the research analyses leading to untrustworthy or misleading empirical findings (1.5. in Table 2)

Arguably the most frequently discussed causes of nonreplicability are *flaws in the research analyses* including HARKing (hypothesizing after the results are known, meaning theories explaining statistically significant results are formulated a posteriori, after looking at empirical data, which is problematic in the context of the hypothetico-deductive research logic and the logic of significance testing), p-hacking (conducting many significance tests until the desired significant result turns up, which leads to misinterpretations of the meaning of the p-value), small sample sizes (leading to imprecise or wrong p-values and effect sizes), misinterpreted p-values, and ignored or not properly considered effect sizes (Bosco et al., 2015; Kerr, 1998; Rubin, 2017; Schönbrodt & Perugini, 2013).

Flaws in the way empirical findings are reported (1.6. in Table 2)

Lacking transparency concerning data collection and data analysis procedures constitute flaws in the way empirical findings are reported (e.g., Bakker & Wicherts, 2011). This includes masquerading exploratory findings as if they were the result of a hypothetico-deductive theory testing process (HARKing, p-hacking), as well as lacking transparency about how findings were obtained (lacking information about sampling selection procedures and about self-selection, lacking open code, lacking information about data cleaning procedures). Such a lack of transparency hinders other researchers from applying the exact same methods in relocation attempts and thus makes successful replications difficult to impossible. In line with previous articles on definitions of trustworthy research, we call this lack of transparency a lack of reproducibility (e.g., Syed, 2021); but to the researcher it can look like a lack of replicability if intransparent reports of research methods prevent a researcher in a replication study from applying the exact same methods as the original study

did, without the researcher knowing about the exact differences between their own methods and the methods of the original study.

Flaws in the interpretation of research findings ("inferential crisis"; 1.7 in Table 2)

A further source of non-replicability and other aspects of lacking trustworthiness is the problem of *flawed interpretation of research findings*. Frequent misinterpretations include the occasions in which a significant p-value is misinterpreted as evidence for the null hypothesis or as evidence for the strength of a correlation or as the size of a group difference (in other words, mixing up the meaning of p-values and effect sizes).

Even correctly conducted data analyses can still be misinterpreted (Greenland et al., 2016). For instance, identifying one theoretical model that fits the data well using structural equation models does not rule out the existence of equivalent or alternative models that fit the data equally well or better. Another misinterpretation consists in theory-method gaps, in which as such sound statistical analyses are interpreted as if they answered theoretical questions that they logically have little to do with (Moeller, 2021; Yarkoni, 2022a). The latter has been called the "inferential crisis" (e.g., Starns et al., 2019; Syed, 2021)

Yet another example of misinterpreted but possibly well-conducted analyses is forking (Gelman & Loken, 2014): If researchers let the data and the outcome of prior analyses determine their next analytical steps or net studies, they implicitly decide which path to take in the universe of possible paths. Finding interesting results of this path does not rule out the possibility that taking another route would have led to different, possibly even opposing insights. "There are many roads to statistical significance, and if data are gathered with no

preconceptions at all, it is obvious that statistical significance can be obtained from pure noise, just by repeatedly performing comparisons, excluding data in different ways, examining different interactions and controlling for different predictors, and so forth." (Gelman & Loken, 2014, p. 461). Arguably, such spurious findings are unlikely to be replicated and even if they are, they may look very different if slightly different methods are used in conceptual replication studies. We also discuss this problem as a possible lack of generalizability of research findings across different research groups in the main text in point 3.2 in Table 2.

Flaws in the research infrastructure leading to a lack of cumulative knowledge building ("normativity crisis"; 1.8. in Table 2)

Problematic publishing policies in journals and problematic incentive systems in universities and funding institutions (Fanelli et al., 2015) contribute to questionable research practices that ultimately fail to cumulate knowledge in systematic ways. Publishing decisions by reviewers and editors have often favored publishing counterintuitive headlines and findings, including precognition (the ability to sense the future; Bem, 2011) over less attention-grabbing topics, which creates incentives for researchers to abandon more solid but less surprising topics in favor of topics that turn heads, but topics that turn heads tend to be those no one expected meaning that no one found particularly probable. The pressure to publish such head-turning studies, and in general to "publish or perish" in order to build a career and get further funding and jobs is arguably one of the problematic incentives in the research community contributing to a prevalence of research findings that seem unlikely to be true and replicable in the first place. What some authors call the grant culture, the pressure to

get as much funding as possible, further contributes to the pressure to publish as many research articles as possible and to appear as successful of a researcher as possible, which can incentivize researchers to favor the quantity and shininess over the quality of their research output. Another example of problems in the infrastructure of the research community is the difficulty to publish null findings (findings failing to falsify the null hypothesis). Up until the replicability debate it tended to be very difficult to publish null findings. This, in combination with the aversion of many journals to publish any replication study at all, arguably contributed to the file drawer problem (relevant findings not getting published). This is turn causes an unchecked publication bias in the form of eventually significant results getting published and contradicting evidence not getting published. Arguments for not publishing replication attempts include the concerns that they lack novelty, as well as that failed replications typically represent nonsignificant findings (called "null findings"). The concern about null findings is that conclusions drawn from them lack the logical certainty that Popper hoped to achieve with falsification. If you falsify the null hypothesis that all swans are white by finding a single black one, you can be certain that not all swans are white. As long as you find no none-white swan to falsify the null hypothesis, you never know whether the null or the alternative hypothesis are ultimately correct. Lundh (2019) has called this the normativity crisis.

Appendix C

Practical solutions aiming for a better understanding of boundary conditions have recently been proposed by several authors. We particularly would like to direct the interested reader to Busse et al. (2017), Bryan et al. (2021), Deffner et al. (2022), Uygun Tunç & Tunç, (2022) and Yarkoni (2022a; 2022) as well as the 38 replies to his discussion of the generalizability crisis, as well as to Czibor et al. (2019). Below, we have linked their suggestions, along with our own, to the following steps in a research cycle:

1) Epistemological approaches and decisions,

2) Theory building,

3) Study planning,

4) Data collection,

5) Data analysis,

6) Interpreting data and making inferences about generalizability,

7) Addressing generalizability of a study's findings in an article's discussion section and proposing directions for future research, and

8) Building upon previous studies in cumulative science.

This is no complete list and we expect more solutions to be proposed, since the debate is still ongoing. Therefore, we have set up an interactive document in which readers can crowdsource further suggestions for practical suggestions of how to detect heterogeneity and how to examine and/or improve generalizability. This interactive document can be found here:

https://docs.google.com/document/d/1cgjKORSEDhwm1dW6HYQSmFGsdx2OPJOX/edit?usp=sharing&ouid=106180454400610418202&rtpof=true&sd=true

1) **Broad suggestions that we could not categorize**

- Theorize and hypothesize about boundary conditions. Systematically test for invariance across theoretically plausible boundary conditions. For instance, instead of simply assuming a finding obtained in a German sample to be universally and nomothetically generalizable, systematically test its generalizability in other countries, other populations (e.g., previously under-sampled minorities) and compare it across sub-groups that could plausibly function differently (e.g., individuals with high versus low emotional stability / neuroticism).
- In the study of boundary conditions, scientists should pay attention to certain real-life contexts (Busse et al., 2017, suggestion 5). This is an overall matter of study planning, data collection, analyses, data interpretation and discussion of boundary conditions.
- It is important that scientists give more attention to the contexts while studying boundary conditions. Scientists exploring boundary conditions should consider real-life contexts. Additionally, they should consider the "respective theory from within those [specific] contexts (Busse et al., 2017, suggestion 6, p. 42). *This is an overall matter of study planning, data collection, analyses, data interpretation and discussion of boundary conditions.*
- Scientists exploring boundary conditions under uncertainty should not only examine the context of interest, but also perform cross-context comparisons. In addition, this type of scientist should handle the theories in trial-and-error mode (Busse et al., 2017, suggestion 7). *This is a matter of data collection and data analysis.*
- Scientists should be aware of the range of the theory they borrow and check it, avoiding out-of-range borrowing. Additionally, they should think about the theory's constructs and relationships focusing on the theoretically most important context descriptions (Busse et al., 2017, suggestion 8). *This is a matter of theory building and*

addressing generalizability of a study's findings in an article's discussion section and how to propose directions for future research.

- Scientists should explore boundary conditions along striking dimensions or at least consider them. Besides other tools that are possible, scholars exploring the boundary conditions of a theory should especially consider the following tools: "a) refining constructs b) amending mediators c) amending moderators as theoretical tools for exploring boundary conditions and widening the applicability of a theory." (Busse et al., 2017, suggestion 9, p.44). *This is a matter of theory building, measurement, data collection data analysis, data interpretation and making inferences about generalizability, addressing generalizability of a study's findings in an article's discussion section and proposing directions for future research, as well as building upon previous studies in cumulative science.*
- Authors should be encouraged by reviewers to consider boundary conditions along prominent dimensions, without being overburdened by the reviewers (Busse et al., 2017, suggestion 10). *This refers to review practices.*
- In order to further develop the research of boundary conditions, it is important to publish insignificant results regarding context factors. Such results provide important information in the field of boundary conditions, as they highlight generalizability with respect to the context dimension tested. Therefore, reviewers and editors should regard "a) clearly delineated boundaries as an indicator of rigor, b) admittedly unexplored boundary conditions as a signal of honesty that is not be judged negatively, c) apparent overstatements of generalizability as an indicator of lacking rigor."(Busse et al., 2017, p.45) (or what to publish how to discuss the generalizability in the discussion section of an article and how to address directions for future research (Busse et al., 2017, suggestion 11, p. 45). *This refers to cumulative science practices and reviewer*

decisions, as well as to the discussion of boundary conditions in discussion sections and directions for future research.

- In the context of research on boundary conditions, reviewers and editors "should be more receptive to discussions of lacking influence of variables in theoretical models" (Busse et al., 2017, suggestion 12, p. 46). This refers to the problem that boundary conditions can never be completely known and are multidimensional, which is why it can be difficult to delineate exactly to which hierarchical combination of factors a theory exactly does, or does not apply. This leads to some fuzziness and uncertainty in determining the scope of a theory. Busse et al. (2017) here make a similar point as we did in our manuscript, namely that some uncertainty remains, but the never complete knowledge gained through boundary condition exploration is still better than gaining no knowledge by not exploring boundary conditions. *It is an epistemological problem, but also a point referring to reviewer practices.*
- *Generalizability licenses* (Pearl, 2018; see Deffner et al., 2022) essentially propose that understanding whether and why findings generalize requires us to understand the causal mechanisms represented by an effect, and their relations to causal mechanisms that may differ between populations. Building on recent work on causal inference, the authors propose that "One key idea [of generalizability licenses] is that generalizability does not depend on the presence of sample differences per se or on raw statistical associations. The conditions that license generalization and comparison with other populations depend on the causal relations between variables and the exact mechanisms by which populations differ" (Deffner et al., 2022, p.2). Deffner et al. (2022, p. 3) propose that "a causal framework for cross-cultural research requires us to state (a) what we want to know, that is, the estimand; (b) a generative model of the evidence, that is, a causal model of how the observed data came into existence; (c) a

generative model of how populations may differ; and (d) a tailored estimation strategy that allows us to learn from data." See also other texts on causal inference by e.g., Bareinboim & Pearl (2016), Lundberg et al. (2021), Pearl & Bareinboim (2014). Such a causal framework can be applied to generalizability across populations (Deffner et al., 2022), but why not also propose it to generalizations across time points, individuals? The authors already propose applying it to questions of measurement invariance, which suggests that it should be equally useful to address construct validity generalizability. Much of the research on causality currently relies on the analysis of between-person variance, which implies that the resulting causal models may not apply, i.e., generalize, to some or even all of the individuals in the sample (e.g., Reitzle, 2013; Moeller, 2021; Molenaar, 2004). A solution may be Bolger et al.'s (2019) proposal of methods analyzing heterogeneity in causal processes by examining within-person variability. See also the works on individual treatment effects (e.g., Montoya et al., 2021; Shalit et al., 2017).

- Appropriately consider generalizability and control, across the lab and the field. Besides the correct statistical inference, researchers also have to balance the generalizability (external validity from experimental setting to real-life contexts) when designing an experiment. Meaning to look at whether the casual relationships continue to hold when e.g. subjects or contexts are modified (Czibor et al., 2019). *This refers to study planning, data collection, discussion of research findings and their generalizability, as well as building on previous studies cumulatively by extending the knowledge on their replicability as well as boundary conditions.*

- Do more field experiments, especially natural field experiments. One threat to

generalizability is the change in people's behavior when being in an experimental set-up. Through natural field experiments this thread can be eliminated (Czibor et al., 2019). *This pertains to study planning and data collection.*

- Integrate lab and field experiments as complementary approaches: Since lab and field experiments can both have advantages and disadvantages, it can be beneficial to use both (Czibor et al., 2019). *This refers to study planning, data collection, data interpretation and cumulative scientific practices.*

- *6-step process of a Lakatos'ian systematic replication framework, proposed by Uygun Tunç & Tunç, 2020):* Step 1: The original study, Step 2: a close replication, Step 3: Conceptual replications testing sets of auxiliary hypotheses about predictors, Step 4: Close replication of step 4, Step 5: Conceptual replications testing sets of auxiliary hypotheses about outcomes, Step 6: Close replication of step 5.

2) Epistemological approaches and decisions

- In the first step, researchers should change their general attitude toward boundary conditions and start to understand boundary conditions as an important core element of research and even as an important research topic itself (Busse et al., 2017, suggestion 1)
- There is a need for a clear terminology and interpretation of the term boundary condition. An understanding of what boundary conditions are and what distinguishes them from other concepts must be created. For the sake of simplification, boundary conditions should be considered a univariate function, therefore, some dimensions of contexts should be eliminated (Busse et al., 2017, suggestion 2).

- Consider that there are both inductive and deductive approaches to solving generalizability problems. The inductive approach proposes to explore boundary conditions and sources of heterogeneity, see e.g., Yarkoni, 2022a; 2022b; Busse et al., 2017; Golino et al., 2022). The deductive approach proposes frameworks for how to test auxiliary hypotheses about boundary conditions in systematic replication studies, see e.g., Uygun Tunç & Tunç, 2020). That both approaches currently co-exist should not distract from the issue that their underlying epistemological rationales may logically differ. It seems that more epistemological debate is needed to solve such differences. We find the distinction of different exploration approaches differing by their starting points from Busse et al. (2017, Table 3) insightful, who distinguish between more theory-driven boundary condition explorations (called inside-out exploration of boundary conditions and outside-in exploration of boundary conditions) and the more data-driven boundary condition explorations (called exploration of boundary condition under uncertainty and serendipitous boundary condition exploration).

3) Theory building

- Earlier there was no clear definition of boundary conditions. Therefore, the topic was not taken as seriously as it should have been taken within research. In the first step, researchers should change their general attitude toward boundary conditions and start to understand boundary conditions as an important core element of research and even as an important research topic itself. (Busse et al., 2017, suggestion 1)
- In the study of boundary conditions, scientists should provide a deep understanding of contexts (Busse et al., 2017, suggestion 5)
- "Increased attentiveness in the hypothesis generation phase to the likely sources of heterogeneity in treatment effects" (Bryan et al. 2021, p.2). "Identifying the

moderators of experimental effects can be a powerful tool for identifying causal mechanisms and its value can be harnessed at multiple stages of the theory-building process" (Bryan et al., 2021, suggestion 1, p. 3).

4) **Planning designing empirical studies**

- In the study of boundary conditions, scientists should pay attention to certain real-life contexts (Busse et al., 2017, suggestion 5).
- They should aim to include theoretically plausible moderators in the data collection.
- They should aim to test the invariance of a prior finding systematically in both samples that are similar to the original study (testing replicability) and in samples including other conditions that may theoretically plausibly represent boundary conditions (e.g., other contexts, such as countries or cultures, other individuals, such as individuals with different socioeconomic status, education, vocational interests and experiences, previously under-sampled minorities, etc.

5) **Data collection**

- "Measure characteristics of samples and research contexts that might contribute to heterogeneity. Scholars are moving towards a kind of data collection that includes the careful conceptualization and measurement of potential moderators at every stage of the research pipeline and builds toward eventual tests of these moderators in generalizable samples (for example, participants randomly selected from a defined population)"Bryan et al. 2021, suggestion 2, p.3).

6) **Data analysis**

- As analytic tools for establishing boundary conditions, Busse et al. (2017) propose using amendment of moderators, refinement of construct, and amendment of mediator. Our work with intensive longitudinal studies has taught us that many other meaningful

statistics can differ between datasets and thus limit the generalizability of conclusions drawn in one to the other. For instance, what if the distribution of a variable differs, being unimodal in one sample and multimodal in the other? What if ergodicity only exists in one but not in the other sample? Invariance should be established not only in regard to the effects addressed in the research question and hypothesis, but also in regard to all the statistics that are needed to obtain that effect, including checks of the assumptions of the test examining the effect (e.g., tests for normal distribution, outliers, homoscedasticity, etc.).

- "The use of new, conservative statistical techniques to identify sources of heterogeneity that might not have been predicted in advance" (Bryan et al., 2021 p. 2). "Of course, the integrity of this approach depends on taking careful measures to avoid over-interpreting chance variation, including pre-registered analysis plans and careful control on multiple hypothesis tests; features that are built into many state-of-the-art statistical techniques" (Bryan et al. 2021, p. 4). "For instance, one study over-sampled schools that were expected to have weaker effects, such as very low-achieving schools that were presumed to lack the resources to benefit from a simple motivational treatment, and very high-achieving schools that were expected not to need an intervention. This gave the study sufficient statistical power to test for interactions" (Bryan et al., 2021, step 3, p. 4).
- *Use Indicators of heterogeneity in meta-analyses*: If possible, examine indicators of heterogeneity in meta-analyses, such as Cochran's Q statistic or the I^2 index. For discussions on how to detect heterogeneity in meta-analyses, see e.g., Lakens, 2022

and Higgins &Thompson, 2002). Lakens (2022) also provides R code facilitating the practical implementation of these recommendations.

- If possible, pool datasets with similar variables from different studies, population samples, and research teams, to compare findings across these different data sources with meta-analytic techniques.
- To establish robustness of a finding in terms of invariance across different, presumably equivalent analytic approaches, consider the following approaches (discussed and compared in Aczel et al., 2021):
    - *multiverse analyses* (Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. 2016. Increasing transparency through a multiverse analysis. Perspectives on Psychological Science 11: 702–712. https://doi.org/10.1177/1745691616658637)
    - *Vibration of effects* (Patel et al.)
    - *Specification curve analysis* (Simonsohn et al., 2020)
- To find out whether a group trend (e.g., an average) generalizes to all or most individuals in a sample, start by checking for multi-modality. Avoid analyzing and interpreting multi-modal distributions as if they were uni-modal. You can identify such multimodal mixture distributions by examining the number of modes with violin plots (which are more informative for multimodal mixture distributions than the more commonly used boxplots, see Figures 7 and 8 in the blog version of Matejka & Fitzmaurice, 2017). After inspecting the distributions visually via violin plots, you can test for multimodality, for instance by using Hartigan's dip test (e.g., with the R package diptest, Maechler, 2021, see also Haslbeck et al., 2022). If you are facing a mixture distribution with multiple modes, be very cautious about the meaning represented by the overall sample mean score, because a mean score requires an uni-

modal distribution in order to be a useful measure of a central tendency (Derrible & Ahmad, 2015; Wirtz & Nachtigall, 1998). Use non-parametric tests to examine multi-modal distributions and aim to describe the distribution of each subgroup's measure of central tendency. If you have a posteriori hypothesis about factors that might explain why you found multiple modes in your sample, test these hypotheses in a new study with a new sample to avoid HARKing and still learn more about these potential distinctly functioning subgroups.

- If you describe any sample distribution, be open to the possibility that there are unknown subgroups and that consequently the overall sample might show a mixture distribution with multiple peaks / modes.
- *Keep in mind that a regression / structural equation model specifying temporal or even causal associations between variables can fail to generalize to some or even all individuals in a sample* (see Reitzle, 2013; for a discussion also Moeller, 2021). You can use within-person analyses in combination with frequency counts to examine how many individuals are described by the bi- and multivariate relations represented by the different paths in a between-person structural equation model.
- Keep in mind that individuals (time points, contexts, etc.) can differ in regard to the measurement models and structural models (to use the language used in structural equation modeling). For instance, a factor structure of a measurement instrument or the relation between two measured constructs can differ between conditions. Do not simply assume invariance, test it.
- *Lacking ergodicity:* In many cases, group-based analyses do not reflect what we would find in the individual units in the group (known as lacking ergodicity or Simpson's paradox). You can detect such lacking ergodicity by examining relevant statistics (e.g., an effect you are interested in, but also measurement models) both within the units of

the group (meaning for instance within individuals) and between. However, to do that, you need to first gather data assessing variance and co-variance within and between the units (here: within and between individuals) of the constructs you are interested in. Only if the data collection captures such variation both within and between, you can then use analyses (e.g., multilevel analyses, scatter plots distinguishing the individual (idiographic) within-person trends and the overall between-person trend. Intensive longitudinal studies typically allow for such disentangling of within-person and between-person variation, because of their repeated measures per person (as long as they also gather data from multiple individuals, in addition to multiple time points. Keep in mind that in the case of lacking ergodicity, your group-based coefficients may fail to generalize to some, many, or even all individuals in your sample. To discover and analytically address lacking ergodicity, see practical solutions proposed by e.g., Golino et al. (2022) and Kievit et al. (2013).

- For some sophisticated analyses such as those often seen in intensive longitudinal studies, *more work is needed to develop methods to determine whether two findings in two samples are sufficiently similar to be considered invariant across samples/time points/individuals/measurement instruments and other sources of variation*. Possibly, this will require some debate about the exact definition of what it means for two findings to be "sufficiently similar".

7) Making interpretations and inferences about generalizability

- Keep in mind that without examining the boundary conditions specified in our Table 2 systematically, it is impossible to say whether invariance of any finding between two studies is due to a lack of replicability in terms of lacking trustworthiness / spurious findings, or due to boundary conditions differing between the two samples or

the methods with which they were obtained. Although it is logically impossible to ever rule out hitherto unknown boundary conditions (hidden moderator problem), at least we can make an educated guess after, but not before, systematically studying these boundary conditions.

8) Addressing generalizability of a study's findings in an article's discussion section and how to propose directions for future research

- Isolating contextual influences: It will become important in future research that context descriptions are made (Busse et al., 2017, suggestion 3).
- Even if a certain uncontrollability of the context description can remain, scientists should try to delineate contexts. The delimitation should be as follows: "a) theoretically meaningful insights from the BC exploration can be expected b) practically relevant context delineations are chosen c) context boundaries are unambiguous" (Busse et al., 2017, suggestion 4, p. 40)
- Scientists should consider perplexity (unexpected findings) as a trigger for the exploration of boundary conditions (Busse et al., 2017, suggestion 5; see also Moeller et al., 2022a)
- *Constraints on generality statements (COG)*: Simons and colleagues (2017) suggest that the discussion sections of empirical papers should include a Constraint on generality statements (COG statement). This statement contains an explanation that identifies the explicit target population of the results. The authors wrote (page 1123): "We propose that the discussion section of all primary research articles specify Constraints on Generality (i.e., a "COG" statement) that identify and justify target populations for the reported findings. Explicitly defining the target populations will

help other researchers to sample from the same populations when conducting a direct replication, and it could encourage follow-up studies that test the boundary conditions of the original finding. Universal adoption of COG statements would change publishing incentives to favor a more cumulative science*"*.

9) Building upon previous studies in cumulative science and review practices

- "Large-scale investment in shared infrastructure to reduce the currently prohibitive cost to individual researchers of collecting data—especially field data—in high-quality generalizable samples" (Bryan et al. 2021, p. 2). "Use of large probability-based samples combined with comprehensive measurement and analysis of moderators" (Bryan et al, 2021, suggestion 4, p. 6).

Appendix D:

A reading list for readers interested in debates about generalizability

This is a list of selected reading recommendations giving interested readers first introductions into the generalizability-related topics described in the headlines below.

Please note that we have set up an interactive document in which readers can crowd-source further recommendations and references trying to improve our understanding of generalizability. This interactive document can be found here: https://docs.google.com/document/d/1dZodTPEHkbAHR58CaExh1dM5BiByYo04/edit?usp=sharing&ouid=106180454400610418202&rtpof=true&sd=true.

1. **Texts about generalizability theory**

Bareinboim, E., & Pearl, J. (2013). A General Algorithm for Deciding Transportability of Experimental Results. *Journal of Causal Inference, 1*(1), 107–134. https://doi.org/10.48550/arXiv.1312.7485

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Briggs, D. C., & Wilson, M. (2007). Generalizability in Item Response Modeling. *Journal of Educational Measurement, 44*(2), 131–155. https://doi.org/10.1111/j.1745-3984.2007.00031.x

Higgins, J., & Thompson, S. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Al-Ubaydli, O., & List, J. A. (2013). On the generalizability of experimental results in eco- nomics: with a response to Camerer. NBER Working Paper No. 19666.

Vivalt, E. (2020). How Much Can We Generalize from Impact Evaluations? *Journal of the European Economic Association*, *18*(6), 3045-3089. https://doi.org/10.1093/jeea/jvaa019

2. **Texts proposing (differing) definitions of replicability and generalizability:**
   2.1 ***Articles on the distinction between direct and conceptual replicability***

Feest, U. (2019). Why replication is overrated. *Philosophy of Science*, *86*(5), 895–905. https://doi.org/10.1086/705451

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: are they important?. *Journal of Social Behavior and Personality*, *5*(4), 41–49.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450

2.2 *Articles on the distinction between local and global generalizability*

Brandtstädter, J. (1985). Individual development in social action contexts: Problems of explanation. In J. R. Nesselroade & A. von Eye (Eds.), *Individual development and social change. Explanatory analysis* (pp. 243-264). Academic Press.

Czibor, E., Jimenez-Gomez, D., & List, J.A. (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal*, *86*(2), 371–432. https://doi.org/10.1002/soej.12392

2.3 *Articles discussing generalizability in relation to construct validation:*

Flake J. K., Luong, R., & Shaw, M. (2021). Addressing a Crisis of Generalizability with Large-Scale Construct Validation. *The Behavioral and brain sciences*, *45*, e14. https://doi.org/10.1017/S0140525X21000376

Shadish, W, Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Yarkoni, T. (2022a). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, E1. 1–37. https://doi.org/10.1017/S0140525X20001685

Please also see the 38 responses to Yarkoni in the same issue of the journal here: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/generalizability-crisis/AD386115BA539A759ACB3093760F4824#related-commentaries

2.4 *Articles discussing generalizability in relation to external validity*

Czibor, E., Jimenez-Gomez, D., & List, J. A. (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal*, *86*(2), 371-432. https://doi.org/10.1002/soej.12392

Shadish, W, Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science, 29*(4), 579–595. https://doi.org/10.1214/14-STS486

3. Articles about the generalizability crisis

Yarkoni, T. (2022a). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, E1.1–37. https://doi.org/10.1017/S0140525X20001685

Yarkoni, T. (2022b). Replies to commentaries on the generalizability crisis. *Behavioral and Brain Sciences, 45*. E40. https://doi.org/10.1017/S0140525X21001758

Please also see the 38 responses to Yarkoni in the same issue of the journal here: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/generalizability-crisis/AD386115BA539A759ACB3093760F4824#related-commentaries

4. **Articles discussing the epistemological foundations guiding our definitions of and decisions about generalizability in relation to boundary conditions**

Adorno, T. W., Albert, H., Dahrendorf, R., Habermas, J., Pilot, H., & Popper, K. R. (Eds.) (1976). *The positivist dispute in German sociology*. Heinemann.

Andersson, G. (2019). Karl Popper und seine Kritiker: Kuhn, Feyerabend und Lakatos [Karl Popper and his critics: Kuhn, Feyerabend and Lakatos]. In G. Franco (ed.), *Handbuch Karl Popper [Handbook Karl Popper] (pp. 717–731).* Springer. https://doi.org/10.1007/978-3-658-16239-9_52

Feest, U. (2019). Why replication is overrated. *Philosophy of Science*, *86*(5), 895–905. https://doi.org/10.1086/705451

Lakatos, I. (1971). History of science and its rational reconstructions. In R. C. Buck, & R. S. Cohen (Eds.), *PSA 1970. Boston studies in the philosophy of science* (Vol. 8, pp. 91–136). Dordrecht. http://doi.org/10.1007/978-94-010-3142-4_7

Lakatos, I. (1978). *The methodology of scientific research programmes.* Cambridge University Press. https://doi.org/10.1017/CBO9780511621123

Lakens, D., Uygun Tunç, D., & Necip Tunç (2022). There is no generalizability crisis. *Behavioral and Brain Sciences*, *45*. E25. https://doi.org/10.1017/s0140525x21000340

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115. https://doi.org/10.1086/288135

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Popper, K. (1935). “Induktionslogik" und “Hypothesenwahrscheinlichkeit” [“Inductive logic and probability of hypotheses”]. *Erkenntnis, 5*, 170-172. https://www.jstor.org/stable/20011753

Popper, K. R. (2017). The Logic of Social Science. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *69*, 215-229. https://doi.org/10.1007/s11577-017-0425-6

Tracy, S. J. (2012). The toxic and mythical combination of a deductive writing logic for inductive qualitative research. *Qualitative Communication Research, 1*(1), 109-141. https://doi.org/10.1525/qcr.2012.1.1.109

Uygun Tunç, D., & Tunç, M. N. (2020, May 13). A Falsificationist Treatment of Auxiliary Hypotheses in Social and Behavioral Sciences: Systematic Replications Framework. Preprint: https://doi.org/10.31234/osf.io/pdm7y

5. **Articles about heterogeneity and the heterogeneity revolution**

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour, 5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Moeller, J. (2021). Averting the next credibility crisis in psychological science: Within-person methods for personalized diagnostics and intervention. *Journal for Person-Oriented Research, 7*(2), 53–77. https://doi.org/10.17505/jpor.2021.23795

6. **Articles about the limitations of the nomothetic paradigm and a possible integration of nomothetic versus idiographic methods**

Beck, E. D., & Jackson, J. J. (2020). Consistency and change in idiographic personality: A longitudinal ESM network study. *Journal of Personality and Social Psychology*, *118*(5), 1080–1100. https://doi.org/10.1037/pspp0000249

Beltz, A. M., & Gates, K. M. (2017). Network mapping with GIMME. Multivariate Behavioral Research, 52(6), 789–804. https://doi.org/10.1080/00273171.2017.1373014

Beltz, A. M., Wright, A. G., Sprague, B. N., & Molenaar, P. C. (2016). Bridging the nomothetic and idiographic approaches to the analysis of clinical data. *Assessment, 23*(4)*,* 447-458. https://doi.org/10.1177/1073191116648209

Lundh, L.-G. (2022) The Central Role of the Concept of Person in Psychological Science, Editorial. *Journal for Person-Oriented Research, 8*(2), 38-42. https://doi.org/10.17505/jpor.2022.24853

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Richters, J. E. (2021). Incredible utility: The lost causes and causal debris of psychological science. *Basic and Applied Social Psychology*, *43*(6), 366–405. https://doi.org/10.1080/01973533.2021.1979003

Robinson, O. C. (2011). The Idiographic / Nomothetic Dichotomy: Tracing Historical Origins of Contemporary Confusions. *History & Philosophy of Psychology, 13*(2), 32–39.

7. **Articles about lacking ergodicity and its implication of limited generalizability of research findings from between-person methods to individuals**

McManus, R. M., Young, L., & Sweetman, J. (2022). Psychology is a Feature of Persons, Not Averages or Distributions: The Group-to-Person Generalizability Problem in Social Cognition Research. Pre-print: https://rmmcmanusblog.files.wordpress.com/2023/01/ampps-22-0036.r1_proof_hi.pdf

Golino, H., Christensen, A. P., & Nesselroade, J. R. (2022, August 2). Towards a psychology of individuals: the ergodicity information index and a bottom-up approach for finding generalizations. *Pre-print PsyArXive:* https://doi.org/10.31234/osf.io/th6rm

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*. *2*(4), 201-218. https://doi.org/10.1207/s15366359mea0204_1

8. **Articles proposing solutions facilitating the understanding of person-, time- and context-specific boundary conditions**

Busse, C., Kach, A. P., & Wagner, S. M. (2017). Boundary conditions: What they are, how to explore them, why we need them, and when to consider them. *Organizational Research Methods, 20*(4), 574-609. https://doi.org/10.1177/1094428116641191

Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A causal framework for cross-cultural generalizability. A*dvances in Methods and Practices in Psychological Science*, *5*(3), https://doi.org/10.1177/25152459221106366

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12(6)*, 1123–1128. https://doi.org/10.1177%2F1745691617708630

Uygun Tunç, D., & Tunç, M. N. (2020, May 13). A Falsificationist Treatment of Auxiliary Hypotheses in Social and Behavioral Sciences: Systematic Replications Framework. Preprint: https://doi.org/10.31234/osf.io/pdm7y