Article

Society for General Psychology
APA DIVISION 1

Review of General Psychology
2022, Vol. 26(2) 241–249
© 2021 The Author(s)
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10892680211015635
journals.sagepub.com/home/rgp
SAGE

Context Dependency as a Predictor of Replicability

Mario Gollwitzer[1] and Johannes Schwabe[1]

Abstract

We scrutinize the argument that unsuccessful replications—and heterogeneous effect sizes more generally—may reflect an underappreciated influence of context characteristics. Notably, while some of these context characteristics may be conceptually irrelevant (as they merely affect psychometric properties of the measured/manipulated variables), others are conceptually relevant as they qualify a theory. Here, we present a conceptual and analytical framework that allows researchers to empirically estimate the extent to which effect size heterogeneity is due to conceptually relevant versus irrelevant context characteristics. According to this framework, contextual characteristics are conceptually relevant when the observed heterogeneity of effect sizes cannot be attributed to psychometric properties. As an illustrative example, we demonstrate that the observed heterogeneity of the "moral typecasting" effect, which had been included in the ManyLabs 2 replication project, is more likely attributable to conceptually relevant rather than irrelevant context characteristics, which suggests that the psychological theory behind this effect may need to be specified. In general, we argue that context dependency should be taken more seriously and treated more carefully by replication research.

Keywords

methodology, replication, hidden moderators, context dependency, social psychology

Psychological science is still dazzled by astoundingly large number of published effects that cannot be replicated—even when these replications used the same materials, the same methods, and a similar setup as the original study (Open Science Collaboration, 2015; Shrout & Rodgers, 2018). One interpretation for the difficulty to replicate an original effect is that it never really existed in the first place (i.e., a "false positive") because it was the result of pure chance, bias, or questionable research practices (e.g., "*p*-hacking"; Bakker et al., 2012; Simmons et al., 2011). A second interpretation for failing to replicate an original effect is that it *did* exist, but the replication attempt was unlikely to find it either because it was underpowered (Etz & Vandekerckhove, 2016; Miller, 2009; Miller & Ulrich, 2016), the "replication success" criterion was underspecified, or the chance for successful replication was low *a priori* (Fiedler & Prager, 2018). Finally, a third interpretation—which is the one that we focus on in this article—is that the effect in an original study did in fact exist, but only under contextual conditions that were present in the original and absent in the replication studies. This interpretation assumes that low replicability rates reflect a low generalizability of empirical findings rather than false positives (e.g., Fabrigar & Wegener, 2016; Gilbert et al., 2016; Stroebe & Strack, 2014).

The argument that nonreplicability may be due to an underestimated influence of context has received both praise and blame. Advocates have noted that embracing the idea of context dependency will pave the way to better theory-building and a more solid specification of the boundary conditions under which a hypothesized effect will or will not be observed (e.g., Pettigrew, 2018). More cautionary voices, on the contrary, have argued that assuming context dependency as a potential reason for nonreplicability leads into an epistemological trap and plays into the hands of problematic scientific practices (e.g., using "context" as a *post hoc* explanation for nonreplication would render any replication effort futile by default; Zwaan et al., 2018; see also Lakatos, 1976; Meehl, 1990).

In this article, we argue that context dependency is a neglected issue at least in some areas of psychological science and may indeed account for the heterogeneity (and, thus, for the nonreplicability) of many effects. More

[1] Department of Psychology, Ludwig-Maximilians-Universität, Munich, Germany

Corresponding Author:
Mario Gollwitzer, Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstrasse 13, 80802 Munich, Germany.
Email: mario.gollwitzer@lmu.de

importantly, we will provide a conceptual and analytical framework that helps researchers decide *a posteriori* whether the observed heterogeneity of a particular effect is more likely due to conceptually relevant or rather to conceptually irrelevant context characteristics. This will hopefully enrich the current replicability debate in psychology by demonstrating the importance of the context dependency argument both for theory-building and for replication science.

Defining "Context"

Arguing that low replicability rates reflect "context dependency" requires that the term "context" is clearly defined and not just used as a vague container concept (e.g., Falleti & Lynch, 2009). Here, referring to Cronbach's (1982) classic UTOS framework, we define "context" as anything that can threaten the generalizability of a finding, including

- Sample characteristics ("**U***nits*" in Cronbach's terms), including socio-demographic, cognitive, or motivational features (e.g., individual reasons for participating in a study) that define a sample and may influence the effect being estimated,
- Psychometric characteristics of the operationalized (measured or manipulated) independent variable(s) ("**T***reatment*"),
- Psychometric characteristics of the operationalized (measured) dependent variable(s) ("**O***utcomes*"),
- Characteristics of the **S***etting* in which a study is conducted, including meso-level characteristics (e.g., features of the location in which the study takes place, interactions between the experimenter and the participant) and macro-level characteristics (e.g., culture-specific norms).

Notably, context characteristics differ in their conceptual relevance: Conceptually relevant context characteristics are those that moderate an effect substantively and, thus, qualify a theory. Drawing on structural equation modeling (SEM) terminology (e.g., Kline, 2015), one could say that *conceptually relevant* context characteristics moderate effects in the *structural* part of the model. In psychology, this applies to context characteristics that represent psychologically meaningful moderators of an effect. By contrast, context characteristics that cannot be substantively interpreted—in SEM terminology, all context characteristics that moderate effects in the *measurement* part of the model—are *conceptually irrelevant*. For instance, measurement instruments that are poorly standardized so that their construct validity or their reliability varies considerably between studies may produce variation in effect size estimates and, thus, increase the probability of an unsuccessful replication of the focal effect (Fabrigar & Wegener, 2016; Hussey & Hughes, 2020; D. J. Stanley & Spence, 2014). Other conceptually irrelevant factors include data preprocessing and/or modeling choices, such as the treatment of outliers or the inclusion of conceptually irrelevant covariates (Ioannidis, 2008; Klau et al., 2020; Patel et al., 2015).

Although both conceptually relevant and irrelevant context characteristics can contribute to effect size heterogeneity and, thus, the nonreplicability of an effect, it is important to distinguish between them because they have to be treated differently by replication research: conceptually relevant context characteristics need to be built into a theory (as boundary conditions or "qualifiers"; see Glöckner & Betsch, 2011). Conceptually irrelevant context characteristics, by contrast, need to be statistically or experimentally controlled (e.g., Petty, 2018). Ideally, it would be possible to specify *a priori* whether a certain contextual characteristic is conceptually relevant versus irrelevant. However, this turns out to be very difficult (Earp & Trafimow, 2015). Therefore, replication science needs a tool to estimate the extent to which an effect depends on conceptually relevant versus irrelevant context characteristics *a posteriori*.

Here, we describe a conceptual and analytical framework that can facilitate such an estimation. Before we do so, let us illustrate the difference between conceptually relevant versus irrelevant context characteristics with the "facial feedback effect"—the finding that activating the *zygomaticus major* (i.e., the "smiling muscle"), for instance, by holding a pen sidewise in one's mouth, makes people find cartoons more amusing than, for instance, holding the pen lengthwise in one's mouth (Strack et al., 1988). The facial feedback effect has twice become famous in social psychology: once when it was originally published and once more about 30 years later, when a registered replication project involving 17 laboratories was unable to replicate it (Wagenmakers et al., 2016). Notably, the replication studies differed in a number of aspects from the original study (see Strack, 2016). One difference was that, in the replication studies, participants were monitored via webcams (which did not even exist in the 1980s). Recent research suggests that exactly this difference can account for the nonreplicability of the original effect and that being monitored is a substantive (and psychologically reasonable) boundary condition of the facial feedback effect (A. A. Marsh et al., 2019; Noah et al., 2018).

This "setting" characteristic (according to Cronbach's UTOS framework) can be regarded *conceptually relevant* because it suggests that the facial feedback effect is diminished when cues of being watched are present, that is, when self-monitoring or self-awareness is high. Thus, self-monitoring is a theoretically meaningful boundary condition that should be included in the theory underlying the facial feedback effect as a "qualifier" (e.g., Glöckner & Betsch, 2011). By contrast, the fact that the cartoons that had been used to measure participants' perceptions of funniness in the 1980s are considered considerably less funny 30 years later—a zeitgeist-related context characteristic—may be considered a methodologically important, yet *conceptually irrelevant*, moderator of the facial feedback effect (Strack, 2016). It challenges the construct validity of the independent variable (Fabrigar & Wegener, 2016; D. J. Stanley & Spence, 2014) and, thus, produces a floor effect that makes the focal effect less likely to be detected; but it does not challenge the generalizability of the effect or its underlying theory (see also Fabrigar et al., 2020).

The facial feedback case shows how replication research can inform us about the robustness of an effect; but, that said, unsuccessful replications should not be prematurely interpreted as clear evidence for a false positive (as some scholars tend to do; e.g., Schimmack, 2020). Rather, they can inspire researchers to investigate boundary conditions for an assumed effect and qualifiers in a psychological theory. From that perspective, it is encouraging to see recent follow-ups on the facial feedback effect, which suggest that not only conceptually relevant (such as self-monitoring or self-awareness; A. A. Marsh et al., 2019; Noah et al., 2018; see also Kaiser & Davey, 2017) but also conceptually irrelevant context characteristics—such as the specific stimuli that are used (Coles et al., 2019a)—moderate the effect and explain why some studies find it and others do not. Such endeavors contribute to a truly cumulative science (e.g., Coles et al., 2019b).

Context Characteristics and Effect Size Heterogeneity

The UTOS framework is a helpful taxonomy of context characteristics, but it remains silent about the *extent* to which nongeneralizability is due to unit, treatment, outcome, or setting characteristics. Replication projects—in particular, "deep" replication projects such as Wagenmakers et al.'s (2016) multisite replication of the facial feedback effect or the various ManyLabs project (see below)—can help elucidate the sources underlying effect size heterogeneity. Importantly, even "direct" or "close" replications (Brandt et al., 2014) differ from an original study with regard to context characteristics such as location, participants, experimenter(s), materials, study setting, and so on (Wong & Steiner, 2018). Thus, it may not be surprising that effect sizes vary considerably—not only between original and replication studies but also between study sites in replication projects (de Boeck & Jeon, 2018; T. D. Stanley et al., 2018). For instance, in the ManyLabs 2 replication project (Klein et al., 2018), half of the effects being tested yielded statistically significant heterogeneity statistics.

Interestingly, the ManyLabs projects seem to speak against the assumption that the heterogeneity of an effect size can be meaningfully explained by generic context characteristics: The ManyLabs 1 project, for instance, found no evidence for a moderating effect of study setting (i.e., laboratory vs. online) or study site/sample country (i.e., U.S. vs. international sample) on focal effect sizes (Klein et al., 2014); ManyLabs 2 looked at culture (i.e., replication studies conducted in "WEIRD" = Western, educated, industrialized, rich, democratic vs. "non-WEIRD" societies) as well as the order in which participants completed the tasks/experiments, and found that at least these two features did not substantially moderate their focal effects (Klein et al., 2018). Finally, ManyLabs 3 looked at time (i.e., when participants completed the experiment during the academic semester), study site, and task order, and found only weak moderator effects (Ebersole et al., 2016). Mirroring these results, replication projects conducted with survey experiments in sociology and political science found that "unit" characteristics, such as whether a survey was conducted in a nationally representative versus a convenience sample collected via Amazon's *Mechanical Turk*, did not produce a considerable amount of heterogeneity in effect sizes (Coppock, 2019; Coppock et al., 2018). However, these approaches have been criticized for being atheoretical and arbitrary and, thus, likely to underestimate true context effects (Fabrigar et al., 2020; Wong & Steiner, 2018).

Supporting the notion that contextual differences may explain (non)replicability, Van Bavel et al. (2016) asked trained coders to rate the context dependency of each of the 100 effects included in the "Reproducibility Project: Psychology" (Open Science Collaboration, 2015) and found that these ratings were indeed associated with the replicability of an effect. Although Van Bavel et al.'s (2016) approach to estimate the influence of context dependency is arguably very rough, and their findings have been challenged methodologically (e.g., Inbar, 2016), their analysis is informative and suggests that context dependency should not be overlooked as a predictor of (non)replicability. Unfortunately, Van Bavel et al.'s (2016) approach failed to differentiate between conceptually relevant versus irrelevant context characteristics, which is, as we have discussed above, an important distinction.

Estimating Conceptually Relevant Versus Irrelevant Context Effects

In this article, we describe an empirical strategy to estimate *a posteriori* how much heterogeneity of observed effect sizes in a replication project is due to conceptually relevant versus irrelevant context characteristics. Building upon previous conceptualizations (Fabrigar et al., 2020; Fabrigar & Wegener, 2016), conceptually *irrelevant* context effects produce heterogeneity in observed effect sizes (across replication studies of the same effect) that can be attributed to differences in construct validity and/or reliability of the measures between study sites. The rest, that is, effect heterogeneity that cannot be accounted for by variability in

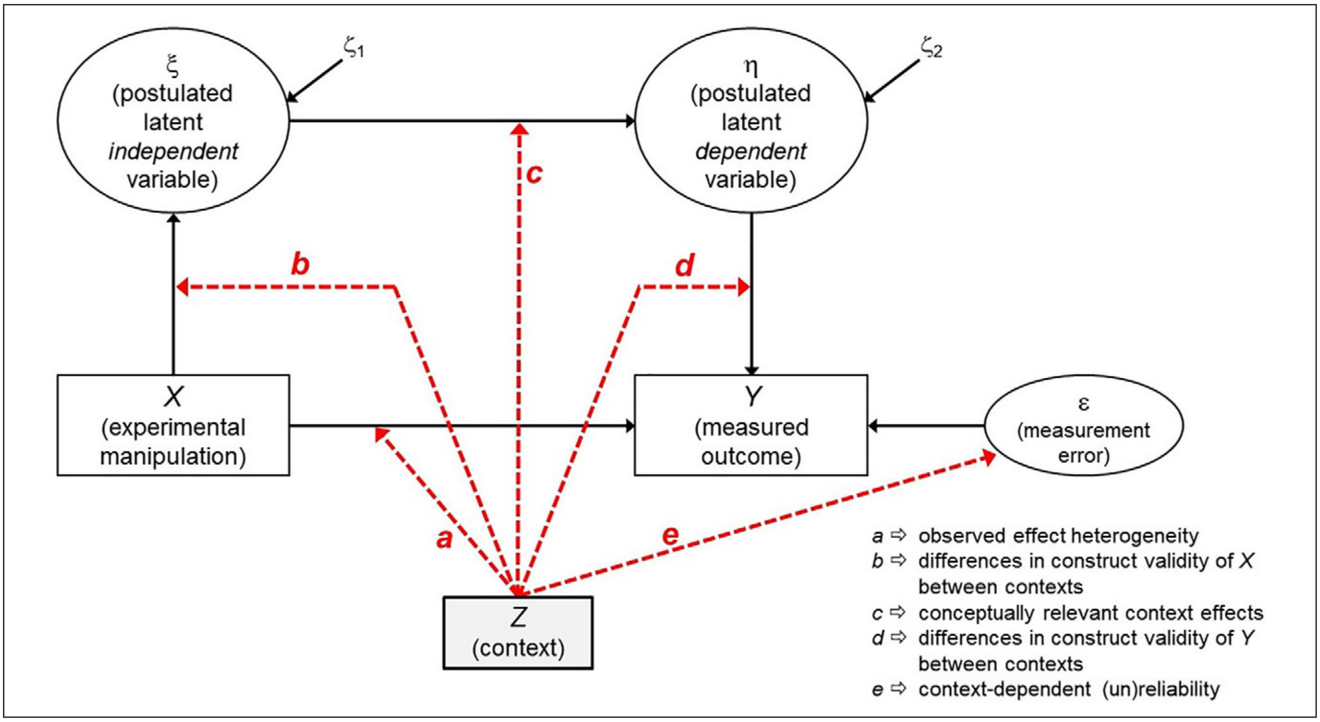
ξ (postulated latent *independent* variable)
ζ_1
η (postulated latent *dependent* variable)
ζ_2
c
b
d
X (experimental manipulation)
Y (measured outcome)
ε (measurement error)
a
e
Z (context)
a ⇨ observed effect heterogeneity
b ⇨ differences in construct validity of X between contexts
c ⇨ conceptually relevant context effects
d ⇨ differences in construct validity of Y between contexts
e ⇨ context-dependent (un)reliability

Figure 1. Structural equation model displaying the effects that "context" (*Z*) can have in an experimental study with one independent and one dependent variable.

psychometric properties of the measures between studies, can be attributed to conceptually *relevant* context effects. In other words, conceptually relevant context effects can be indirectly estimated by subtracting the amount of conceptually irrelevant context effects from the observed heterogeneity of an effect.

Let us take one effect from the ManyLabs 2 project (Klein et al., 2018) as an example: the "moral typecasting" effect originally reported by Gray and Wegner (2009: Study 1a). In this experiment, participants read a story about a person ("offender") who did something that harmed another person ("victim"). The moral typecasting effect says that the amount of responsibility ascribed to the offender (i.e., the dependent variable η) is a function of the offender's agency (i.e., the independent variable ξ). The latent dependent variable "responsibility" (η) was measured with one item (Y_1): "How responsible is Sam for his behavior?" (on a scale from 1 = *not at all responsible* to 7 = *fully responsible*). Two secondary dependent variables (Y_2 and Y_3) asked about the offender's intentionality and the amount of pain felt by the victim. In the original experiment and in the replication studies, the latent independent variable "agency" (ξ) was manipulated by varying the offender's age (*X*; the offense was committed either by an adult man [x_1] or a child [x_2]).

Figure 1 displays the resulting structural equation model (SEM) as a path diagram (using the typical notation from SEM; see Kline, 2015): the path from *X* to ξ signifies that varying *X* (i.e., age) aims to cause variation in the theoretically assumed independent variable ξ (i.e., agency); and the path from η to *Y* signifies that *Y* is a manifestation of the latent dependent variable η (responsibility). Context characteristics—that is, all UTOS characteristics that may differ systematically between study sites—are denoted as *Z* here.

In the replication of the "moral typecasting" effect (ManyLabs 2; Klein et al., 2018), researchers at all 61 study sites used the original materials from Gray and Wegner (2009), which may suggest little context variation between study sites. However, the replication project was conducted in 27 countries, and the materials had to be translated into 12 different languages (which may cause systematic variation in "treatment" and/or "outcome" characteristics). Thirty-eight labs recruited their participants from a pool, 23 did not. At three sites, participants completed the study in a classroom, 31 sites used a lab setting, and at 25 sites, participants completed the study online at home (i.e., systematic variation in "setting"). Also, sample characteristics varied across study sites: for instance, the relative frequency of male participants varied between 1% and 75% (i.e., systematic variation in "units").

By no means, we intend to suggest that these differences render this replication project useless or the data uninterpretable—on the contrary. But researchers should keep in mind that such differences may well produce effect size heterogeneity that may be either conceptually relevant or irrelevant (Wong & Steiner, 2018). More specifically, these context differences (*Z*) can produce heterogeneity in the observed $X \rightarrow Y$ effect between studies (denoted as "*a*" in Figure 1). This heterogeneity is a function of (1) the extent to which the manipulation "worked" across different contexts ("*b*"), (2) the "true" context dependency of the effect ("*c*"), (3) measurement (in)variance regarding the dependent variable ("*d*"), (4) measurement error (i.e., unreliability in *Y*), which may also differ between contexts ("*e*"), and (5) sampling error (which we will not discuss further here). To distill *c*, the variability caused by *b, d*, and *e* need to be estimated and subtracted from the observed heterogeneity.

To illustrate how this can be achieved, we will again use the moral typecasting effect (Gray & Wegner, 2009; see also the description of how this effect was replicated in the ManyLabs 2 project by Klein et al., 2018, p. 464) as an example. Note however, that the direct replication approach used in ManyLabs 2 necessarily leads to an underestimation of both method as well as true heterogeneity (McShane et al., 2016). Thus, the following should be understood as a proof of concept, and not as an actual attempt to attain optimal estimates.

The moral typecasting effect was tested at 61 different sites with sample sizes ranging between $n = 16$ and $n = 841$ per site. The effect size in the original study was $d = 0.80$; the effect size obtained in the replication project (i.e., computed across all participants, irrespective of study site) was slightly higher ($d = 0.95$); the median effect size across study sites was $d = 1.04$. Thus, the effect can be considered successfully replicated. Nevertheless, there was quite some heterogeneity across study sites (Klein et al., 2018; Table 3). Another way to look at this heterogeneity is to analyze the data in a multilevel model with two levels (participants nested within study sites; Snijders & Bosker, 2012). Analyzing the data that way corroborates the successful replication of the focal effect, fixed effect of experimental manipulation: $B = 0.783$, $SE(B) = 0.037$, 95% CI for *B* [0.709, 0.858], but also clearly shows that the effects varied considerably across study sites, as reflected by a significant random slope variance $\hat{\sigma}^2_{U_1} = 0.057$, $SE(\hat{\sigma}^2_{U_1}) = 0.015$, 95% CI for $\hat{\sigma}^2_{U_1}$ [0.033, 0.097]. Does this heterogeneity point to conceptually relevant or rather conceptually irrelevant context conditions?

First, the extent to which the experimental manipulation "worked" invariantly at the 61 different study sites (context effect *b* in Figure 1) could play a role, although it is reasonable to assume that the experimental manipulation used by Gray and Wegner (2009: adult vs. child offender) is so explicit and blatant that it is likely to have worked well and similarly at each study site, so one might safely assume that $b = 0$ here. For subtler manipulations, however, it is useful to investigate how strongly the effect of the manipulation varies across context, for instance, by conducting a careful manipulation check (Fabrigar et al., 2020; Fiedler et al., 2021).

Second, the assumption of strict measurement invariance across study sites might be violated (context effects *d* and *e* in Figure 1). If more than one indicator (i.e., item) per latent variable are used in a study, then these effects can be estimated via multigroup confirmatory factor analysis (Meredith, 1993) or the *alignment* method (Asparouhov & Muthén, 2014; H. W. Marsh et al., 2018), which is particularly suited for a large number of groups. Here, we specified a very simple measurement model (assuming that the "responsibility" item and the "intentionality" item used by Gray & Wegner, 2009, reflect two manifestations of the same latent factor) and compared item loadings and error variances between the 27 countries in which this study was replicated. With this model, it is possible to scrutinize the level of heterogeneity in factor loadings and error variances across groups (here: countries) and to determine those groups that contribute most strongly to a violation of the measurement invariance assumption. Here, we found no evidence for such a violation: neither factor loadings nor error variances varied substantially between countries. Thus, context effects *d* and *e* (see Figure 1) can also be ruled out, and the heterogeneity across study sites that was observed in this replication project is most likely due to a "true" context dependency regarding the effect of agency on responsibility ratings (i.e., context effect *c*). Thus, the next step would then be to look for these conceptually relevant context characteristics, describe and explain them, and amend the theory accordingly.

Discussion

This illustrative example shows that it is possible to estimate how much variability in observed effect sizes in a replication project is due to conceptually irrelevant versus conceptually relevant context characteristics. The conceptual and analytical framework proposed here allows researchers to decide *a posteriori* whether exploring conceptually relevant context characteristics is worth doing, and whether the theory underlying the proposed effect needs to be specified and enriched with "qualifiers" (Glöckner & Betsch, 2011), which, at least for the moral typecasting effect, seems to be the case.

Notably, context characteristics are not the same as confounders (and should not be treated as such). Confounders—measured or hidden factors that are related both to the independent (*X*) as well as the dependent variable (*Y*)—artificially inflate or deflate the true association between *X* and *Y*. Thus, confounders may produce either false-positive or false-negative results, which are nonreplicable when the replication studies are void of the respective confounders. Context characteristics, by contrast, do not bias the estimated association between *X* and *Y*; rather, they represent "true" moderator effects and cause heterogeneity in observed effect sizes.

Importantly, some context characteristics that may appear conceptually irrelevant on the surface may turn out to be conceptually relevant upon second thought: for instance, effect size heterogeneity between different cultures may reflect either measurement invariance (i.e., a conceptually irrelevant context characteristic) or a psychologically meaningful and, thus, conceptually relevant, boundary condition of the focal effect (Fabrigar et al., 2020). Moreover, some context characteristics reflect both conceptually relevant *and* irrelevant conditions at the same time. For instance, Karau and Williams (1993) meta-analytically showed that "social loafing" effects—the finding that the size of a group is inversely related to the individual amount of effort invested to the group task by each individual member—are larger in student samples than in working adult samples. This "unit" characteristic (according to Cronbach's UTOS framework) may reflect both a conceptually relevant condition (i.e., social loafing may be more easy to detect in professional team contexts than in student collaboration projects, and detectability is a substantive contextual moderator of social loafing) and a conceptually irrelevant condition (e.g., student samples are more homogeneous and, thus, produce smaller sampling errors and larger effects; see Peterson, 2001, but see Hanel & Vione, 2016).

Determining whether a context characteristic reflects a conceptually relevant or rather an irrelevant condition may sometimes be challenging, but it is also necessary and important. Thus, specifying a theory with regard to its conceptually relevant conditions should be an ongoing, iterative process that requires a systematic and collaborative effort, and replication research offers a tool to undertake this effort (e.g., Asendorpf et al., 2013; see also Coles et al., 2019b). Observing that an effect varies significantly between studies and that this variation is not due to methodological factors alone should prompt a search for conceptual moderators or boundary conditions. In addition, heterogeneity analysis may be used to decide whether or not a conceptual moderator or a boundary condition should be formally included in the current state of a theory. Ideally, theory databases would provide researchers with pertinent information about the current state of a theory, including all known conceptual moderators and boundary condition as well as information on the degree of empirical corroboration of each single proposition (e.g., in form of one constantly updated Bayes-Factor based on all studies testing the same conceptual hypothesis; see Glöckner & Betsch, 2011; Glöckner et al., 2018).

The framework presented here rests on the assumption that observed heterogeneity can be reliably estimated. This requires that a large-scale replication project (such as "ManyLabs") has already been conducted. Notably, meta-analyses should not be used to estimate observed heterogeneity for the presently described purpose because—contrary to a replication project—the original studies being included in a meta-analysis have been selected post hoc, and, in most cases, these original studies had never been designed to estimate the heterogeneity of the focal effect in the first place (Purgato & Adams, 2012). This can lead to biases (McShane et al., 2016), which are further amplified by publication biases such as journal editors' reluctance to publish nonsignificant results (LeBel et al., 2018; Zwaan et al., 2018). Therefore, estimating the impact of context dependency ideally requires a representative number of replication studies that cover as many dimensions of "context" as possible (for example, using a "meta-studies" framework, see Baribault et al., 2018).

In addition, improved estimates for the impact of "treatment" and "outcome" characteristics (i.e., effects *b, d*, and *e* in Figure 1) are needed. Ideally, information about the measurement properties of a manipulated or measured variable could be drawn easily and directly from a database consisting of the raw data of each study in which the respective variable has been measured and validated (Hussey & Hughes, 2020). While the idea of such a measurement-specific raw database may have sounded unrealistic and visionary 10 years ago, the "open science" movement, which is a central element in "psychology's renaissance" (Nelson et al., 2018; Nosek et al., 2012) has shown that it is possible to set up such a database.

Conclusion

Determining the extent to which observed effect size heterogeneity reflects conceptually relevant versus irrelevant context characteristics is important because a number of relevant implications can be derived from such an analysis. First, if a true effect is small and varies strongly across different contexts, the results of single replication studies are barely more (or less) informative than original studies (Kenny & Judd, 2019; McShane et al., 2019). Second, the statistical power to detect a small, yet variable effect should rather be maximized by increasing the number of studies than by increasing the sample size of each individual study (Kenny & Judd, 2019). Third, knowledge about the sources of heterogeneity is relevant for theory development and theory specification. Heterogeneity being due to conceptually relevant context characteristics suggests that these context characteristics need to be built into the theory as qualifiers or boundary conditions. Thus, elucidating the heterogeneity of psychological effects is highly informative not only to enrich the replicability debate but for psychological science as a whole.

Acknowledgments

The authors thank Joachim I. Krueger and John Rauthmann for helpful comments on an earlier version of this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by an "LMUexcellent" investment grant.

ORCID iD

Mario Gollwitzer https://orcid.org/0000-0003-4310-4793

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119. https://doi.org/10.1002/per.1919

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 495–508. https://doi.org/10.1080/10705511.2014.919210

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. https://doi.org/10.1177/1745691612459060

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., de Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 2607–2612. https://doi.org/10.1073/pnas.1708285114

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Coles, N. A., Larsen, J. T., & Lench, H. C. (2019a). A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychological Bulletin*, *145*(6), 610–651. https://doi.org/10.1037/bul0000194

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Arinze, N. C., Ndukaihe, I. L. G., Özdoğru, A. A., Aczel, B., Hajdu, N., Nagy, T., Som, B., Basnight-Brown, D., Zambrano, D., Javela, L. G., Foroni, F., Willis, M., Pfuhl, G., Kaminski, G., Ehrengarth, T., IJzerman, H., . . . Liuzza, M. (2019b). *A multi-lab test of the facial feedback hypothesis by the Many Smiles Collaboration*. https://doi.org/10.31234/osf.io/cvpuw

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, *7*, 613–628. https://doi.org/10.1017/psrm.2018.10

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 12441–12446. https://doi.org/10.1073/pnas.1808083115

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.

de Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin*, *144*, 757–777. https://doi.org/10.1037/bul0000154

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, Article e621. https://doi.org/10.3389/fpsyg.2015.00621

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., & Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLOS ONE*, *11*, Article e0149794. https://doi.org/10.1371/journal.pone.0149794

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80. https://doi.org/10.1016/j.jesp.2015.07.009

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, *24*, 316–344. https://doi.org/10.1177/1088868320931366

Falleti, T. G., & Lynch, J. F. (2009). Context and causal mechanisms in political analysis. *Comparative Political Studies*, *42*, 1143–1166. https://doi.org/10.1177/0010414009331724

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science*. Advance Online Publication. https://doi.org/10.1177/1745691620970602

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—Illustrated by the report of the Open Science Collaboration. *Basic and Applied Social Psychology*, *40*, 115–124. https://doi.org/10.1080/01973533.2017.1421953

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science." *Science*, *351*, Article e1037. https://doi.org/10.1126/science.aad7243

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making*, *6*, 711–721.

Glöckner, A., Fiedler, S., & Renkewitz, F. (2018). Belastbare und effiziente Wissenschaft [Durable and efficient science]. *Psychologische Rundschau*, *69*, 22–36. https://doi.org/10.1026/0033-3042/a000384

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*, 505–520. https://doi.org/10.1037/a0013748

Hanel, P. H. P., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? *PLOS ONE*, *11*, Article e0168354. https://doi.org/10.1371/journal.pone.0168354

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*, 166–184. https://doi.org/10.1177/2515245919882903

Inbar, Y. (2016). Association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, e4933–e4934. https://doi.org/10.1073/pnas.1608676113

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Kaiser, J., & Davey, G. C. L. (2017). The effect of facial feedback on the evaluation of statements describing everyday situations and the role of awareness. *Consciousness and Cognition*, *53*, 23–30. https://doi.org/10.1016/j.concog.2017.05.006

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*, 681–706. https://doi.org/10.1037/0022-3514.65.4.681

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*, 578–589. https://doi.org/10.1037/met0000209

Klau, S., Schönbrodt, F., Patel, C., Ioannidis, J., Boulesteix, A.-L., & Hoffmann, S. (2020). *Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology* (Technical Reports, Nr. 232). Department of Statistics, Ludwig-Maximilians-Universität München. https://doi.org/10.5282/UBM/EPUB.70485

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability—A “many labs” replication project. *Social Psychology*, *45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., & Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. https://doi.org/10.1177/2515245918810225

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes [Monographs on Epistemology, Logic, Methodology, Philosophy of Science, Sociology of Science and of Knowledge, and on the Mathematical Methods of Social and Behavioral Sciences]. In S. G. Harding (Ed.), *Can theories be refuted? Synthese Library* (Vol. 81, pp. 205–259). Springer. http://doi.org/10.1007/978-94-010-1863-0_14

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*, 389–402. https://doi.org/10.1177/2515245918787489

Marsh, A. A., Rhoads, S. A., & Ryan, R. M. (2019). A multi-semester classroom demonstration yields evidence in support of the facial feedback effect. *Emotion*, *19*, 1500–1504. https://doi.org/10.1037/emo0000532

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*, *23*, 524–545. https://doi.org/10.1037/met0000113

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730–749. https://doi.org/10.1177/1745691616662243

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician*, *73*, 99–105. https://doi.org/10.1080/00031305.2018.1505655

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

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. https://doi.org/10.1007/BF02294825

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*, 617–640. https://doi.org/10.3758/PBR.16.4.617

Miller, J., & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science*, *11*, 664–691. https://doi.org/10.1177/1745691616649170

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology’s renaissance. *Annual Review of Psychology*, *69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology*, *114*, 657–664. https://doi.org/10.1037/pspa0000121

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. https://doi.org/10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

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*, 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*, 450–461. https://doi.org/10.1086/323732

Pettigrew, T. F. (2018). The emergence of contextual social psychology. *Personality & Social Psychology Bulletin*, *44*, 963–971. https://doi.org/10.1177/0146167218756033

Petty, R. E. (2018). The importance of exact conceptual replications. *Behavioral and Brain Sciences*, *41*, Article e146. https://doi.org/10.1017/S0140525X18000821

Purgato, M., & Adams, C. E. (2012). Heterogeneity: The issue of apples, oranges and fruit pie. *Epidemiology and Psychiatric Sciences*, *21*, 27–29. https://doi.org/10.1017/s2045796011000643

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie canadienne*, *61*, 364–376. https://doi.org/10.1037/cap0000246

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE.

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*, 305–318. https://doi.org/10.1177/1745691614528518

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*, 1325–1346. https://doi.org/10.1037/bul0000169

Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science*, *11*, 929–930. https://doi.org/10.1177/1745691616674460

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*, 768–777. https://doi.org/10.1037/0022-3514.54.5.768

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

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 6454–6459. https://doi.org/10.1073/pnas.1521897113

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkstra, K., & Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*, 917–928. https://doi.org/10.1177/1745691616674458

Wong, V. C., & Steiner, P. M. (2018). *Replication designs for causal inference* [Working Paper]. https://curry.virginia.edu/sites/default/files/uploads/epw/62_Replication_Designs.pdf

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, Article e120. https://doi.org/10.1017/S0140525X17001972