


Looking our limitations in the eye: A call for more thorough and honest reporting of study limitations

Beth Clarke¹  | Lindsay J. Alley² | Sakshi Ghai³ |
Jessica K. Flake² | Julia M. Rohrer⁴ | Joseph P. Simmons⁵ |
Sarah R. Schiavone¹ | Simine Vazire¹

¹Melbourne School of Psychological Sciences, University of Melbourne, Parkville, Victoria, Australia

²Department of Psychology, McGill University, Montreal, Quebec, Canada

³Oxford Internet Institute, University of Oxford, Oxford, UK

⁴Wilhelm Wundt Institute for Psychology, Leipzig University, Leipzig, Germany

⁵Department of Operations, Information, and Decisions, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence

Beth Clarke.

Email: bethc1@student.unimelb.edu.au

Funding information

Australian Government Research Training Program Scholarship

Abstract

The replication crisis and subsequent credibility revolution in psychology have highlighted many suboptimal research practices such as *p*-hacking, overgeneralizing, and a lack of transparency. These practices may have been employed reflexively but upon reflection, they are hard to defend. We suggest that current practices for reporting and discussing study limitations are another example of an area where there is much room for improvement. In this article, we call for more rigorous reporting of study limitations in social and personality psychology articles, and we offer advice for how to do this. We recommend that authors consider what the best argument is against their conclusions (which we call the “steel-person principle”). We consider limitations as threats to construct, internal, external, and statistical conclusion validity (Shadish et al., 2002), and offer some examples for better practice reporting of common study limitations. Our advice has its own limitations – both our representation of current practices and our recommendations are largely based on our own metaresearch and opinions. Nevertheless, we hope that we can prompt researchers to write more deeply and clearly about the

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). Social and Personality Psychology Compass published by John Wiley & Sons Ltd.

limitations of their research, and to hold each other to higher standards when reviewing each other's work.

KEYWORDS

limitations, research methods and practices, the four validities, writing

1 | INTRODUCTION

Over the course of psychology's replication crisis and subsequent credibility revolution, our field has come face to face with several suboptimal habits many of us had adopted, often without thinking deeply about the practices we were engaging in. For example, many of us treated Hypothesizing After the Results are Known (HARKing; Kerr, 1998) as an acceptable practice and even taught it to our students (Vazire, 2014). The credibility revolution has prompted a great deal of self-reflection and widespread bottom-up efforts to improve our norms and practices (for a review of some of these reforms see Nelson et al., 2018 and Nosek et al., 2022). In this article, we draw attention to another common but suboptimal practice in psychology publications: the cursory and hand-waving reporting of study limitations.

Limitations are an inherent part of the research process. When limitations are honestly and accurately reported, they offer a more complete and balanced picture of the research at hand. For instance, if limitations are not reported accurately (and conclusions are not calibrated appropriately) this could lead consumers of research to have misplaced confidence in the research, to apply research before it is ready for application, or to misapply findings to real-world contexts (e.g., policy decisions) in domains that are beyond the scope of the research. The field's attitudes towards limitations can have more far-reaching consequences too—if the field were to gain a reputation for overhyping claims and underselling limitations, such a reputation would likely erode trust in our field (Hoekstra & Vazire, 2021; Ramsey, 2021). Discussing study limitations is just one small part of communicating science responsibly (see Corneille et al., 2023) and ensuring that our research claims are calibrated, but improving this practice would help bolster our field's credibility.

The standard practice for reporting limitations in social and personality psychology leaves much room for improvement. Concerningly, only 62% of short articles in social and personality psychology mentioned a limitation (Clarke et al., 2023). In longer format articles, the rate of reporting limitations may be slightly higher (between 63% and 89% of articles in the fields of management, industrial-organizational psychology, entrepreneurship and leadership, Brutus et al., 2013), however, since there is no other empirical meta-research on the prevalence of reported limitations in social and personality psychology articles, it is unclear whether these rates are higher because of differences in the length of the articles examined, or because there are differences between fields. In our experience, even when limitations are acknowledged, there is a lack of discussion regarding the implications those limitations have for the study's conclusions. Instead, authors often note a limitation and then proceed to either: (a) justify the limitation, for example, on the basis of limited resources, or by stating that the limitation is common in the field; or, (b) dismiss the importance of the limitation (particularly by discussing how the limitation is actually a strength; a similar sentiment is echoed by Lingard, 2015).

In psychology there are not strong norms regarding how limitations should be handled. While the American Psychological Association publication manual provides some advice about the kinds of limitations authors should consider when interpreting their results (e.g., “sources of potential bias and other threats to internal validity,” “the imprecision of measures,” “the adequacy of sample sizes and sampling validity” etc. APA, 2020, p. 90), when it comes to writing the final manuscript, implementing this advice is not so straightforward.

In this article, we call for psychology researchers to more rigorously discuss their studies' limitations in empirical articles. We limit our recommendations to the writing stage of research, after the research has been carried out. However, not all limitations can be addressed at the writing stage. Limitations that fundamentally threaten or undermine the validity of the claims could be grounds for abandoning the project altogether. For example, if scores from measures are invalid, or if the statistical tests are known to be misleadingly biased, acknowledging these as limitations is (in most cases) insufficient; instead, fatal limitations must be considered when designing the study and planning the analyses. For the present purposes, our advice is directed towards non-fatal limitations.

Our advice is directly relevant to authors, who, having conducted the research, will often be best placed to acknowledge their studies' limitations. However, we also call on reviewers and editors, who play a crucial role in influencing how authors discuss their studies' limitations and conclusions (Hoekstra & Vazire, 2021; Lingard, 2015; Ramsey, 2021). Indeed, the perception that reviewers could use limitations as ammunition may make authors reluctant to report them. The extent to which this perception matches reality is unclear, but in any case, editors and reviewers should be mindful that their practices are not accidentally rewarding authors who hide or downplay limitations. We hope this article can help reviewers and editors use their power for good, by incentivizing more rigorous discussion of limitations and holding authors accountable for un- or under-reported limitations.

1.1 | Channeling your inner steel-person

Ultimately, the purpose of reporting study limitations is to inform readers about potential threats to the validity of the conclusions, to discuss how these threats might impact the strength of the evidence in supporting those conclusions, and to identify boundary conditions. To better ensure limitations fulfill this purpose, authors can take the perspective of an exceptional critic when discussing their studies' limitations. What would a highly competent but extremely skeptical reader say about your study? By providing the strongest possible case for their study's limitations, authors can "steel-person" their limitations which is to refute the strongest possible opposing argument (i.e., the opposite of the more commonly known practice of using a "strawman/person" argument). Here, we suggest some guiding principles for channeling your inner steel person:

1. Focus on your most important limitations. Throwing a large number of potentially trivial limitations at readers may distract them from the more important points. Instead, highlight your biggest and worst limitations. Consider which limitations an exceptional critic would say are the most damning. Focus on doing these limitations justice, rather than trying to cover as many limitations as possible.¹
2. Be specific. Avoid discussing limitations in a way that is generic and uninformative (e.g., "Due to the observational design of our study, no causal inferences can be drawn"). This is not to suggest that you should refrain from reporting common limitations; limitations that are widespread and common in the field are important to discuss if they impact what conclusions can be drawn. However, the discussion of those limitations should contextualize them in terms of how the limitations threaten the conclusions you draw.
3. Explain, don't excuse. Avoid the urge to justify your study's limitations, to downplay their significance, or to paint limitations as strengths. Instead, elaborate on your study's limitations by explaining their implications for your conclusions. These may include: increased risk of bias, increased uncertainty, alternative explanations that cannot be ruled out, and more. The more specific you can be about implications (e.g., discussing the direction and magnitude of bias, as recommended by the STROBE guidelines, von Elm et al., 2007) the better.
4. Don't seek refuge in future directions. It may be difficult to dwell on your limitations, and future directions provide a tempting opportunity to redirect attention away from your ugly limitations and towards rosy future studies. Resist this urge. Instead, first focus on what the limitations mean for the current study. Then, if there is a fix that is not obvious and would be helpful for researchers planning follow-up studies, explain it and be specific.

But assume your readers already know that a more reliable measure, a more representative sample, or better experimental control, would be better.

5. Do not leave your limitations in exile. Limitations entail caveats; it is imperative that these caveats are integrated in all conclusions throughout the article, including those presented in abstracts. Reporting a limitation *somewhere* does not absolve you of having to take the limitation into account throughout the rest of the article. Sometimes important limitations can even be incorporated into titles, for example, to acknowledge the population (e.g., "... in U.S. college students"), the design (e.g., "a cross-sectional study of..."), or the measure used (e.g., "... predicts self-reported health").

In short, limitations should have clear and concrete consequences for authors' interpretations. In the remaining sections, we offer examples for how authors could discuss a few common limitations in social and personality psychology (see Tables 1–8). We structure these examples according to the four validities framework (the four validities being construct validity, internal validity, external validity, and statistical conclusion validity; Shadish et al., 2002; also see Albright & Malloy, 2000; Kenny, 2019; Vazire et al., 2022). For readers that may be unfamiliar with this framework, we define each of the four validities in turn below. For each of the four validities, we discuss only two common limitations to illustrate how current practices could be improved; these examples are by no means exhaustive of all validity threats or limitations (for an extended list of threats to validity, visit seaboat.io; Schiavone et al., 2023).

Of course, the four validities framework has its own limitations. For instance, although we discuss them in isolation, the four validities are interdependent. For example, valid causal inferences (i.e., strong internal validity) require valid interpretations of measures and manipulations (i.e., strong construct validity), well-specified samples and contexts (i.e., strong external validity), and statistical models that are interpreted appropriately (i.e., strong statistical conclusion validity). Nevertheless, this framework is useful when thinking through and communicating research limitations.

2 | CONSTRUCT VALIDITY

Construct validity is about understanding, manipulating, and measuring constructs (Shadish et al., 2002). In supporting their rationale for a particular use of a measure, it is recommended that researchers collect many sources of construct validity evidence (content, response processes, internal structure, relations to other variables, and consequences of testing; for more details see American Educational Research Association [AERA] et al., 2014). Additionally, as validation is an ongoing process, the fact that a measure or manipulation has previous validity evidence or is frequently used does not necessarily mean that it is valid for use in the study at hand (AERA, 2014). Phrases such as "the scale has been validated" should be avoided. Authors should be more specific: report what validity evidence exists; explain how the evidence does or does not support the authors' use; and, most importantly, discuss what evidence is lacking. We discuss two specific limitations related to construct validity: overreliance on quantitative evidence and the use of on-the-fly measures and manipulations.

2.1 | Over-reliance on quantitative evidence

Some types of validity evidence are reported much more frequently than others. Researchers typically rely on the internal structure of the measure and its relations to other variables, and neglect response processes and consequences of testing (Cizek et al., 2008; Hogan & Agnello, 2004; Villalobos Coronel, 2015). Additionally, many studies report a reliability coefficient as the only "validity" evidence (Flake et al., 2017) but this is not actually an indication of validity (Schmitt, 1996). No number can tell you how participants understand items, or whether item content

aligns with the construct definition. Considering only quantitative evidence often results in maximizing reliability at the expense of validity (Clifton, 2020). If only quantitative evidence (such as a factor analysis or reliability) are reported there could still be important reasons to worry about the validity of the measures and these should be discussed as limitations.

2.2 | On-the-fly measurement and manipulation

Measures and manipulations are often developed by researchers on-the-fly for use in a specific study (i.e., ad hoc or impromptu measures). Previous reviews have found that from 47% (Flake et al., 2017) to 69% (Weidman et al., 2017) of measures and 81% of manipulations (Chester & Lasko, 2021) examined were developed this way. While in some cases validity evidence is collected as part of the study, the majority of ad hoc measures and manipulations rely on no evidence beyond face validity (Chester & Lasko, 2021; Weidman et al., 2017). Given this, it is notable that the use of on-the-fly measures was rarely mentioned as a limitation in a sample of short-report articles examined (Clarke et al., 2023). When on-the-fly manipulations are used, validity evidence is sometimes collected in the form of a pilot study and/or manipulation check (Chester & Lasko, 2021). However, these are often assessed with measures that have unknown or variable validity evidence themselves. Whenever an on-the-fly measure or manipulation is used, and especially when little or no validity evidence is collected, this should be discussed as an important limitation and authors' conclusions should be appropriately caveated.

3 | INTERNAL VALIDITY

Internal validity is about the validity of claims regarding what mechanisms underlie causal relationships (i.e., causality). Threats to internal validity are often discussed in a generic fashion, with authors stating that the study was observational, or that X was not manipulated, and that, therefore, causal conclusions cannot be made (although often causal claims are still made; Grosz et al., 2020). Readers will usually be aware of the general limitations, so the specifics are more interesting: How could inferences be biased in this study; how plausible are such biases? The specific threats will largely depend on the study's design. While experiments are often considered the "gold standard" for providing causal evidence, they rely on some assumptions and are subject to their own limitations (Diener et al., 2022). We discuss two specific reasons why strong causal conclusions may not be warranted: unobserved confounding and measured (not manipulated) mediators.

TABLE 1 Limitation threatening construct validity (example): Over-reliance on quantitative evidence of a measure's validity.

Example of limitation	The "love of books" measure was developed entirely on the basis of quantitative evidence (e.g., factor analysis, reliability).
Suboptimal practice	No limitation mentioned.
Better practice	"For our measure of love of books, we do not have evidence regarding how respondents interpret items or formulate their responses, and whether this is consistent with our construct definition – we only have evidence of the measure's reliability and factor structure. When developing the measure, we chose to retain items with high factor loadings, resulting in a smaller, more narrow set of items than might be ideal. As such, our instrument may not measure the full breadth of the construct 'love of books'. Our incomplete evidence regarding the measure's validity means that scores on this measure may reflect something other than love of books, or a narrow conceptualization of it." Incorporate this limitation in all conclusions throughout the article.
Recommended reading	"Measure twice, cut down error: A process for enhancing the validity of survey scales." (Gehlbach & Brinkworth, 2011)

TABLE 2 Limitation threatening construct validity (example): On-the-fly measures/manipulations.

Example of limitation	The “love of books” measure was developed on-the-fly for use in the study, and no validity evidence was collected.
Suboptimal practice	No limitation mentioned.
Better practice	“Our measure of love of books was developed for use in this study, and no validity evidence was collected. Thus, the validity of this measure is unknown, and we interpret our results as preliminary.” Ensure all conclusions are consistent with this limitation.
Recommended reading	“Constructing validity: New developments in creating objective measuring instruments.” (Clark & Watson, 2019)

TABLE 3 Limitation threatening internal validity (example): Unobserved confounding.

Example of limitation	Testing whether breastfeeding increases childrens' intelligence using an observational design. The study includes only an imprecise measure of a likely confound—socio-economic status.
Suboptimal practice	“This study was merely observational and thus does not justify causal conclusions. Future experimental studies...”
Better practice	“Because socio-economic status likely impacts both breastfeeding and child intelligence, it is an important confound that must be considered when reflecting on the relationship between breastfeeding and child intelligence. We could only control for crude measures of socio-economic status, thus our estimates are most likely biased upwards.” Incorporate this limitation in all conclusions throughout the article.
Recommended reading	“Statistical control requires causal justification” (Wysocki et al., 2022)

3.1 | Unobserved confounding

When conducting non-experimental studies, the universe of potential unobserved confounders (i.e., problematic “third variables”) is infinite. Discussion of these threats should focus on those that are most plausible and potentially have large effects on X and Y. Importantly, authors should identify at least a few central confounders that cannot be ruled out by the study at hand (e.g., socio-demographic factors, personality variables). Often one can add an educated guess about the direction of the potential bias. Furthermore, when analyzed appropriately, observational longitudinal data can to some degree rule out unobserved time-invariant confounders - but not time-varying confounders (Rohrer & Murayama, 2023). Even potential confounders that were measured and accounted for could still be threats if the scores from measures are invalid or unreliable.

3.2 | Mediation analysis

Mediation analyses that investigate the effects of X on Y via mediator M are often conducted on cross-sectional data, although this practice is generally discouraged (Pek & Hoyle, 2016). Another common type of study investigates mediation in a scenario in which X but not M has been manipulated. Here, the internal validity of the effect of the manipulation on the mediator is usually unproblematic (so long as the manipulation and measures are valid, there is no selective attrition, etc.). However, because M has not been manipulated, the internal validity of the indirect effect (which is usually the effect of interest) and of the remaining direct effect can be threatened (Bullock, et al., 2010; MacKinnon & Pirlott, 2015; Rohrer et al., 2022). In particular, the possibilities of reverse causation (Y causing M), of unobserved confounding between the mediator and the outcome (a third variable causing both M

TABLE 4 Limitation threatening internal validity (example): Common mediation analysis practice.

Example of limitation	Testing whether the effects of a brief intervention on school grades is mediated by grit. Grit has been measured, not manipulated.
Suboptimal practice	No limitation mentioned.
Better practice	“In this experimental study, we found that a brief intervention boosted students' school grades. The effect of the intervention may be partially mediated by grit, although any factors that affect both grit and school grades in the same direction (e.g., wealth) may lead to an overestimation of the indirect effect.” Incorporate this limitation in all conclusions throughout the article.
Recommended reading	“Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis.” (MacKinnon & Pirlott, 2015)

and Y), and of collider bias distorting the direct effect (see Rohrer et al., 2022 for an explanation of collider bias) should be considered.

4 | EXTERNAL VALIDITY

External validity is the validity of claims about how the results would hold with other populations, settings, operationalizations, etc. Psychology's lack of sample diversity is one of the most common limitations mentioned, both in articles themselves (Clarke et al., 2023) and in broader discourse about the field (e.g., Henrich et al., 2010; Puthillam et al., 2024). Another important limitation related to external validity is the degree to which study designs create artificial contexts that limit the authors' ability to draw conclusions about the real-world phenomena they are aiming to understand or explain. We consider these two limitations (lack of sample diversity and low realism) below.

4.1 | Lack of sample diversity

Representative samples are rare in psychology (Rad et al., 2018). The specific gaps between the sample and the target population, and the threats that these gaps present, are different for every study and should be considered in the context of each study's aims and designs. Authors should carefully consider the ways in which a lack of sample diversity is relevant to the conclusions they wish to make, and should consider making their target population explicit by incorporating a “Constraints on Generality” section (Simons et al., 2017). For instance, samples rarely represent groups that are marginalized and vulnerable within the study's cultural context (Ghai et al., 2023; Medin, 2017). This is true both within and outside of Western contexts. A thoughtful reflection on the sample's composition (e.g., geographic origin; participant demographics) and likely selection biases (if the sample was not representative) will help inform the reader about the extent of a study's generalizability (see Simons et al., 2017 for guidance on writing a Constraints on Generality section; also see Rad et al., 2018 for an alternative suggestion on explicitly tying findings to the population).

Even large, diverse samples do not necessarily lead to better external validity if they are biased towards (or against) groups that are unusual on the variables of interest (Nagler & Tucker, 2015). Thus, authors should caveat how their samples are biased in known ways (e.g., highly-educated, more digitally-connected), flag uncertainty about unknown sources of bias, and discuss how this limits the generalizability of their conclusions. Importantly, such a discussion should go beyond simply describing the likely biases, and should discuss the consequences of these biases for interpreting results.

4.2 | Low realism

Many studies in social and personality psychology are conducted in contexts that are quite different from the rich real-world phenomena that the studies aim to shed light on (Anderson et al., 2019). This is an understandable compromise. For example, researchers may conduct studies in the laboratory in order to have more experimental control, or they may study a phenomenon using materials that can be administered online in order to recruit a larger or more diverse sample (indeed, online samples are becoming increasingly common in social and personality psychology, Anderson et al., 2019). However, this compromise threatens conclusions about how the findings relate to the real-world phenomenon of interest, and this threat should be taken seriously.

In experimental studies, the manipulated variable(s)'s operationalization may require diverging from the conceptual causal variable as it naturally occurs (Rozin, 2001). As a result, a finding may not hold beyond a study's selected stimuli, tasks, and procedures (Holleman et al., 2020). When this is a serious possibility, authors should circumscribe their conclusions to settings and operationalizations that closely resemble those of their study, and refrain from drawing conclusions or making recommendations about real world contexts that differ from that of the study.

TABLE 5 Limitation threatening external validity (example): Lack of sample diversity.

Example of limitation	Using a convenience sample to test the effects of social comparison on body dissatisfaction.
Suboptimal practice	"Our sample was drawn from an undergraduate population. Future research should test the generalizability of this finding."
Better practice	"Our sample was culturally homogenous, lacking diverse representation of people both within and outside the US. Specifically, most participants were female students, caucasian, born between the mid-1990s to early-2000's, and recruited from a university in Texas. This demographic may be especially sensitive to social comparison with regard to body dissatisfaction, compared to other genders and other age groups (Myers & Crowther, 2009). Thus, the negative association reported here is likely to be inflated due to these selection biases." Incorporate this limitation in all conclusions throughout the article.
Recommended reading	"Constraints on Generality (COG): A proposed addition to all empirical papers" (Simons et al., 2017)

TABLE 6 Limitation threatening external validity (example): Low realism.

Example of limitation	Testing individual differences in moral judgments in an online study using hypothetical vignettes (e.g., the trolley problem).
Suboptimal practice	"Participants responded to hypothetical vignettes about moral judgments. The extent to which these vignettes elicit responses that reflect the same mechanisms as responses to real-world situations is unclear."
Better practice	"We studied moral judgments using trolley problem vignettes, which have been criticized for several reasons (see Bauman et al., 2014). Most importantly, trolley problem vignettes may evoke different psychological processes than moral decision making in the real world. For instance, there is some evidence to suggest participants often find trolley problem scenarios to be humorous, which could lead them to pay less attention to the central moral dilemma in the vignette to maintain their mood (Bauman et al., 2014). Our findings are therefore preliminary and offer suggestive evidence for relationships that must be probed further before they can be applied in the real world." Incorporate this limitation in all conclusions throughout the article.
Recommended reading	"External validity" (Findley et al., 2021)

Although observational studies can often be conducted in more varied and natural contexts, they often still have to adopt artificial design features. For example, using simplified stimuli or hypothetical scenarios can undermine external validity. It is important for observational researchers to question the realism of their measures and clearly specify the boundaries of their findings (Findley et al., 2021). When there is a significant threat to realism, authors should caution readers that their conclusions are not ready to be applied to real-world settings.

5 | STATISTICAL CONCLUSION VALIDITY

Statistical conclusion validity is the validity of statistical inferences, including the risk of errors, bias in estimated effect sizes, the appropriateness of statistical tests, whether assumptions are met, and more. The often fatal nature of threats to statistical conclusion validity may explain why they were the rarest category of limitation in the short-report articles examined by Clarke et al. (2023). However, there are some situations in which threats to statistical conclusion validity are survivable (but still not immediately fixable), and in those cases, limitations should catalog these threats and their implications for the conclusions. We discuss two such limitations: flexibility in data analysis and weak statistical evidence.

5.1 | Flexibility in data analysis

Preregistration is still far from the norm in psychology (Hardwicke et al., 2022 found that only 3% of psychology articles published between 2014 and 2017 provided a preregistration). Research results that emerge from unplanned analyses—analyses that are not preregistered or that meaningfully deviate from preregistered plans—are exploratory. This is, of course, OK, as long as conclusions are appropriately calibrated—many research findings that emerge from exploratory investigations turn out to be true and useful.

Nevertheless, conducting many unplanned analyses usually increases the chance that the significant ones turn out to be false-positives (you are more likely to roll at least one 7 on a 20-sided die if you roll it multiple times). It is hard for researchers to keep track of the number of exploratory analyses they have run; it is impossible for readers

TABLE 7 Limitation threatening statistical conclusion validity (example): Conflating exploratory and confirmatory research.

Example of limitation	Confirmatory statistical tools are used (e.g., Null Hypothesis Significance Testing) to test the effectiveness of a wellbeing intervention, but analyses are not preregistered or decisions about data collection or analysis were not constrained by the preregistration.
Suboptimal practice	No limitation mentioned.
Better practice	“While we found evidence that our intervention improved wellbeing, the analyses testing this effect were not constrained by a detailed preregistered analysis plan [OR: <i>we deviated from our preregistered analysis plan, introducing flexibility</i>]. This increases the chances that these findings are false positives. Thus, these findings are preliminary and, like most exploratory findings, should be considered tentative unless and until stronger evidence emerges.” Incorporate this limitation in all conclusions throughout the article.
Recommended reading	“Pre-registration: Why and how” (Simmons et al., 2021)

of a study to know how many additional analyses were conducted. Thus, a finding from an exploratory analysis is more likely to be a false-positive.

In an ideal world, researchers would publish articles that include preregistered replication studies of novel findings, particularly where the results hinge on unplanned analyses (Simmons et al., 2021), but the world is not ideal, and practical constraints in the form of time, money, or access to the relevant participant population may prevent researchers from carrying out those replications in a timely fashion. In that case, researchers should report their exploratory findings while labeling them as such, and emphasizing that the results are tentative. This can also be emphasized in titles.

Relatedly, researchers who deviate from their preregistrations, or who subsequently realize that their preregistrations were not specific enough to describe exactly how the analyses would be conducted, should explicitly say so in the Method and Results sections, as well as when discussing their limitations. Most importantly, this should reduce confidence in the results (to a greater or lesser degree depending on the amount and nature of the flexibility), and the certainty with which authors interpret those results. If researchers choose to conduct unplanned analyses, it should be clear from the paper that these analyses were not preregistered and the results from these analyses should be interpreted as exploratory.

5.2 | Weak statistical evidence

When the results of a study do not provide much evidence to guide statistical conclusions, it is important to acknowledge weak statistical evidence as a limitation. For example, in the context of Null Hypothesis Significance Testing, p -values between 0.01 and 0.05 are considered by some to be merely suggestive, especially in the context of non-preregistered research (Benjamin et al., 2018). Other examples of weak statistical evidence include: results with 95% confidence intervals that are very wide, and/or come close to including values that are inconsistent with the research hypothesis; precisely-estimated effect sizes that are so small that they could be considered trivial; Bayes Factors between 1 and 3 (or between 1 and $\frac{1}{3}$; Stefan et al., 2019); statistical results that are not robust to alternative ways of analyzing the data; a set of p -values, across multiple studies, that are not heavily right-skewed (for strong statistical evidence, most p -values should be close to zero, Simonsohn et al., 2014).

Weak statistical evidence is rarely, if ever, acknowledged as a limitation—but it should be. If, for instance, the main finding is supported by a few p -values in the ballpark of 0.02–0.10, this should be flagged as weak statistical evidence (see Simonsohn et al., 2014, for an explanation), and researchers should emphasize that in their interpretation of their findings. And in their Abstract. And in their communications with the New York Times.

TABLE 8 Limitation threatening statistical conclusion validity (example): Weak statistical evidence.

Example of limitation	The key finding, that priming participants to think about the future makes them more impulsive, was supported by p -values that were barely statistically significant.
Suboptimal practice	No limitation mentioned.
Better practice	“Our key finding—that priming participants to think about the future makes them more impulsive—was supported by results that were barely statistically significant, and thus should be considered tentative.” Incorporate this limitation in all conclusions throughout the article.
Recommended reading	“P-curve: A key to the file-drawer.” (Simonsohn et al., 2014)

6 | CONCLUSION

In our experience (having read many articles in the field and studied the limitations reported in them; Clarke et al., 2023) our field could do more to truly grapple with our limitations. We believe that other fields face similar problems—we have focused on social and personality psychology because it is our field of expertise and not because we think it is particularly bad. We call on other researchers to also reflect on the state of reported limitations in their fields (as has already been done in several fields including management Brutus et al., 2013, nursing, Connelly, 2013, and medicine, Ross & Zaidi, 2019).

Here we tackle one part of that problem by offering guidance to authors, reviewers, and editors that will enable more thorough discussion of research limitations. Our examples, categorized into threats to the four validities, are illustrative, not exhaustive. We hope they provide a blueprint for engaging more deeply with limitations beyond the examples discussed.

Naturally, our article also has important limitations. First, our scope is limited—our advice does not cover all types of research conducted in social and personality psychology. Second, our portrayal of current practices and our advice are largely based on our opinions, which are informed through our metaresearch, but are opinions nonetheless. We expect many researchers will disagree with some of our characterizations of common practice, or advice about best practice, and we may certainly be wrong. We would be thrilled if this article spurs lively debate about best practices for reporting limitations.

Admittedly, the current incentive structure largely discourages the humility we are calling for. Researchers may fear that an honest self-examination of limitations—and incorporating them fully into their conclusions—may make their work seem less impressive. However, it is unclear how accurate these perceptions are. In our roles as authors, readers, reviewers, and editors, we contribute to these incentives and we can therefore choose to encourage and reward authors who really grapple with their limitations. If evaluators are competent and critical, a more honest account of one's research limitations should earn authors more, not less, credibility (holding the actual limitations of the research constant). Authors could also consider Registered Reports as an avenue that enables a more honest and thorough reporting of study limitations in the Discussion section, because it is written after In Principle Acceptance.²

The flaws in our field's research methods and practices have been a source of turmoil over the course of the recent replication crisis. Our first priority should be to improve our research practices so that our studies have fewer serious limitations. However, it would be overly idealistic to suggest that we could, even with our best efforts, eliminate all serious limitations. If we grapple with these limitations more rigorously, limitations can be an opportunity to demonstrate a commitment to accuracy and rigor.

AUTHOR CONTRIBUTIONS

Beth Clarke: Conceptualization; project administration; resources; writing – original draft; writing – reviewing & editing. **Lindsay J. Alley:** Conceptualization; writing – original draft; writing – reviewing & editing. **Sakshi Ghai:** Conceptualization; writing – original draft; writing – reviewing & editing. **Jessica K. Flake:** Conceptualization; writing – reviewing & editing. **Julia M. Rohrer:** Conceptualization; writing – original draft; writing – reviewing & editing. **Joseph Simmons:** Conceptualization; writing – original draft. **Sarah R. Schiavone:** Conceptualization; writing – reviewing & editing. **Simine Vazire:** Conceptualization; project administration; supervision; writing – original draft; writing – reviewing & editing.

ACKNOWLEDGMENTS

Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST STATEMENT

None.

ORCID

Beth Clarke  <https://orcid.org/0000-0002-7355-6718>

ENDNOTES

¹ It is particularly important that short format papers focus on only the most important limitations given there will be limited words to discuss those limitations. To ensure the limited word count is used most efficiently, authors should also consider principles 3 and 4 by refraining from excusing limitations, and from focusing on future directions.

² We thank a reviewer for raising this suggestion in their review.

REFERENCES

- Albright, L., & Malloy, T. E. (2000). Experimental validity: Brunswik, campbell, cronbach, and enduring issues. *Review of General Psychology*, 4(4), 337–353. <https://doi.org/10.1037/1089-2680.4.4.337>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- American Psychological Association. (2020). *Publication manual of the* (7th ed.). American Psychological Association. <https://doi.org/10.1037/0000165-000>
- Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, 45(6), 842–850. <https://doi.org/10.1177/0146167218798821>
- Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology: External validity in moral psychology. *Social and Personality Psychology Compass*, 8(9), 536–554. <https://doi.org/10.1111/spc3.12131>
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ..., & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10. <https://doi.org/10.1038/s41562-017-0189-z>
- Brutus, S., Aguinis, H., & Wassmer, U. (2013). Self-reported limitations and future directions in scholarly reports: Analysis and recommendations. *Journal of Management*, 39(1), 48–75. <https://doi.org/10.1177/0149206312455245>
- Bullock, B., John, Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, 98(4), 550–558. <https://doi.org/10.1037/a0018933>
- Chester, D. S., & Lasko, E. N. (2021). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspectives on Psychological Science*, 16(2), 377–395. <https://doi.org/10.1177/1745691620950684>
- Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68(3), 397–412. <https://doi.org/10.1177/0013164407310130>
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412–1427. <https://doi.org/10.1037/pas0000626>
- Clarke, B., Schiavone, S. R., & Vazire, S. (2023). What limitations are reported in short articles in social and personality psychology? *Journal of Personality and Social Psychology*, 125(4), 874–901. <https://doi.org/10.1037/pspp0000458>
- Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, 25(3), 259–270. <https://doi.org/10.1037/met0000236>
- Connelly, L. M. (2013). Limitation section. *Medsurg Nursing*, 22(5), 325.
- Corneille, O., Havemann, J., Henderson, E. L., Ijzerman, H., Hussey, I., Orban De Xivry, J.-J., Jussim, L., Holmes, N. P., Pilacinski, A., Beffara, B., Carroll, H., Outa, N. O., Lush, P., & Lotter, L. D. (2023). Beware 'persuasive communication devices' when writing and reading scientific articles. *Elife*, 12, e88654. <https://doi.org/10.7554/eLife.88654>
- Diener, E., Northcott, R., Zyphur, M. J., & West, S. G. (2022). Beyond experiments. *Perspectives on Psychological Science*, 17(4), 1101–1119. <https://doi.org/10.1177/17456916211037670>
- Findley, M. G., Kikuta, K., & Denly, M. (2021). External validity. *Annual Review of Political Science*, 24(1), 365–393. <https://doi.org/10.1146/annurev-polisci-041719-102556>

- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, 8(4), 370–378. <https://doi.org/10.1177/1948550617693063>
- Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, 15(4), 380–387.
- Ghai, S., Fassi, L., Awadh, F., & Orben, A. (2023). Lack of sample diversity in research on adolescent depression and social media use: A scoping review and meta-analysis. *Clinical Psychological Science*. <https://doi.org/10.31234/osf.io/s7juz>
- Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, 15(5), 1243–1255. <https://doi.org/10.1177/1745691620921521>
- Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, 17(1), 239–251. <https://doi.org/10.1177/1745691620979806>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. Cambridge Core. <https://doi.org/10.1017/S0140525X0999152X>
- Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature Human Behaviour*, 5(12), 1602–1607. <https://doi.org/10.1038/s41562-021-01203-8>
- Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement*, 64(5), 802–812. <https://doi.org/10.1177/0013164404264120>
- Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The ‘real-world approach’ and its problems: A critique of the term ecological validity. *Frontiers in Psychology*, 11, 721. <https://doi.org/10.3389/fpsyg.2020.00721>
- Kenny, D. A. (2019). Enhancing validity in psychological research. *American Psychologist*, 74(9), 1018–1028. <https://doi.org/10.1037/amp0000531>
- 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
- Lingard, L. (2015). The art of limitations. *Perspectives on Medical Education*, 4(3), 136–137. <https://doi.org/10.1007/s40037-015-0181-0>
- MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review*, 19(1), 30–43. <https://doi.org/10.1177/1088868314542878>
- Medin, D. L. (2017). Psychological science as a complex system: Report card. *Perspectives on Psychological Science*, 12(4), 669–674. <https://doi.org/10.1177/1745691616687746>
- Myers, T. A., & Crowther, J. H. (2009). Social comparison as a predictor of body dissatisfaction: A meta-analytic review. *Journal of Abnormal Psychology*, 118(4), 683–698. <https://doi.org/10.1037/a0016763>
- Nagler, J., & Tucker, J. A. (2015). Drawing inferences and testing theories with big data. *PS: Political Science & Politics*, 48(01), 84–88. <https://doi.org/10.1017/S1049096514001796>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology’s renaissance. *Annual Review of Psychology*, 69(1), 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- 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>
- Pek, J., & Hoyle, R. H. (2016). On the (In)Validity of tests of simple mediation: Threats and solutions. *Social and Personality Psychology Compass*, 10(3), 150–163. <https://doi.org/10.1111/spc3.12237>
- Puthillam, A., Montilla Doble, L. J., Delos Santos, J. J. I., Elsherif, M. M., Steltenpohl, C. N., Moreau, D., Pownall, M., Silverstein, P., Anand-Vembar, S., & Kapoor, H. (2024). Guidelines to improve internationalization in the psychological sciences. *Social and Personality Psychology Compass*, 18(1), e12847. <https://doi.org/10.1111/spc3.12847>
- Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, 115(45), 11401–11405. <https://doi.org/10.1073/pnas.1721165115>
- Ramsey, R. (2021). A call for greater modesty in psychology and cognitive neuroscience. *Collabra: Psychology*, 7(1). <https://doi.org/10.1525/collabra.24091>
- Rohrer, J. M., Hünemann, P., Arslan, R. C., & Elson, M. (2022). That’s a lot to process! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science*, 5(2), 251524592210958. <https://doi.org/10.1177/25152459221095827>
- Rohrer, J. M., & Murayama, K. (2023). These are not the effects you are looking for: Causality and the within-/between-persons distinction in longitudinal data analysis. *Advances in Methods and Practices in Psychological Science*, 6(1), 251524592211408. <https://doi.org/10.1177/25152459221140842>

- Ross, P. T., & Zaidi, N. L. B. (2019). Limited by our limitations. *Perspectives on Medical Education*, 8(4), 261–264. <https://doi.org/10.1007/s40037-019-00530-x>
- Rozin, P. (2001). Social psychology and science: Some lessons from solomon asch. *Personality and Social Psychology Review*, 5(1), 2–14. https://doi.org/10.1207/S15327957PSPR0501_1
- Schiavone, S. R., Quinn, K. A., & Vazire, S. (2023). A consensus-based tool for evaluating threats to the validity of empirical research [preprint]. <https://doi.org/10.31234/osf.io/fc8v3>. PsyArXiv.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350–353. <https://doi.org/10.1037/1040-3590.8.4.350>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Simmons, J., Nelson, L., & Simonsohn, U. (2021). Pre-registration: Why and how. *Journal of Consumer Psychology*, 31(1), 151–162. <https://doi.org/10.1002/jcpy.1208>
- 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/10.1177/1745691617708630>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-Curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534–547. <https://doi.org/10.1037/a0033242>
- Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, 51(3), 1042–1058. <https://doi.org/10.3758/s13428-018-01189-8>
- Vazire, S. (2014). Life after bem. *Sometimes i'm Wrong*. <https://sometimesimwrong.typepad.com/wrong/2014/03/life-after-bem.html>
- Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, 31(2), 162–168. <https://doi.org/10.1177/09637214211067779>
- Villalobos Coronel, M. (2015). Synthesis of reliability and validation practices used with the Rosenberg self-esteem scale. <https://doi.org/10.14288/1.0165784>
- von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P., & for the STROBE Initiative. (2007). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Annals of Internal Medicine*, 147(8), 573. <https://doi.org/10.7326/0003-4819-147-8-200710160-00010>
- Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, 17(2), 267–295. <https://doi.org/10.1037/emo0000226>
- Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, 5(2), 251524592210958. <https://doi.org/10.1177/25152459221095823>

How to cite this article: Clarke, B., Alley, L. J., Ghai, S., Flake, J. K., Rohrer, J. M., Simmons, J. P., Schiavone, S. R., & Vazire, S. (2024). Looking our limitations in the eye: A call for more thorough and honest reporting of study limitations. *Social and Personality Psychology Compass*, e12979. <https://doi.org/10.1111/spc3.12979>