Construct validity and the validity of replication studies: A systematic review

Jessica Kay Flake

McGill University

Ian J. Davidson

Concordia University of Edmonton

Octavia Wong

York University

Jolynn Pek

The Ohio State University

Accepted for publication at American Psychologist on March 18, 2022
 The final article will be available, upon publication, via its DOI: 10.1037/amp0001006

Author Contributions: JKF, ID, and OW coded the data. JKF and OW analyzed the data. JP, JKF, and ID designed the study. All authors wrote the manuscript and contributed to quality checking the data and analyses.

Acknowledgements: We would like to thank and acknowledge the research assistants that helped to double code the data: Bianca Uguccioni, Andrew Kim, and Jun Choi. This research was funded by the SSHRC small grants program (P2016-0202) and the Early Researcher Award granted by the Ontario Ministry of Research and Innovation (ER15-11-004), both awarded to Jolynn Pek.

Open Science: All RPP materials and our final coding sheet, detailed materials, analysis code and output are hosted on the Open Science Framework (RPP materials: https://osf.io/ezcuj/wiki/home/; our review materials https://osf.io/k3qdn/).

Correspondence concerning this article should be addressed to Jessica Kay Flake, Department of Psychology, McGill University, Montreal QC. kayflake@gmail.com

Abstract

Currently there is little guidance for navigating measurement challenges that threaten construct validity in replication research. To identify common challenges and ultimately strengthen replication research, we conducted a systematic review of the measures used in the 100 original and replication studies from the Reproducibility Project Psychology (Open Science Collaboration, 2015). Results indicate that it was common for scales used in the original studies to have little or no validity evidence. Our systematic review demonstrates and corroborates evidence that issues of construct validity are sorely neglected in original and replicated research. We identify four measurement challenges replicators are likely to face: a lack of essential measurement information, a lack of validity evidence, measurement differences, and translation. Next, we offer solutions for addressing these challenges that will improve measurement practices in original and replication research. Finally, we close with a discussion of the need to develop measurement methodologies for the next generation of replication research.

Keywords: Measurement, psychometrics, open science, replication, methodological reform

Public Significance:

Over the past decade psychologists have been calling for methodological reform to increase the rigor and replicability of psychological science, which has been accompanied by progress in improving transparency and statistical practices. This paper presents rigorous measurement practices as foundational for generating knowledge from psychological science that can be translated to inform policy, develop interventions, and improve people's lives. We review one of the largest sets of replication studies ever conducted to understand how measurement can be improved and develop measurement practices for the next generation of replication research.

Introduction

Psychology is more than a decade into actively evaluating the replicability of study conclusions and rethinking methodological practices (Open Science Collaboration, 2015; Simmons et al., 2011). Scholars have argued for the value of direct replication, a study in which researchers repeat a previous study as closely as possible by using the same procedures, methods, and analysis as a means of evaluating the published literature (Plucker & Makel, 2021; Simons, 2014). Large-scale replication studies involving groups of researchers are now common, ongoing, and highly cited (e.g., Jones et al., 2021; Klein et al., 2014). Following suit, editorial practices have shifted to encouraging replication research with explicit support for direct replications (Simons et al., 2014) and new avenues for publishing in respected journals (e.g., registered replication reports in *Advances in Methods and Practices in Psychological Science*).

This is welcome reform; the replicability of results is at the core of scientific progress. When scientists see the same pattern across multiple studies theories are corroborated, discoveries are made, and breakthroughs occur. And when the same pattern does not emerge, scientists are inspired to revise theory and method to gain further understanding. Although a replication study may initially seem straightforward to implement, repeating the same study to see if you get the same result can be complex. The purpose of this paper is to demonstrate that construct validity evidence is neglected in replication research and ignoring construct validity gives rise to replication results that are uninterpretable or invalid. First, we explain what construct validity is; then to understand how construct validity has been evaluated in replication research, we conducted a systematic review of the measures used in the 100 original and replication studies from the Reproducibility Project: Psychology (RPP; Open Science Collaboration, 2015). Harnessing the observations from our review, we present

recommendations on how to address the invalidity of original and replication studies and discuss measurement practices for the next generation of replication research.

Construct Validity

When a number represents someone's level of motivation, an aspect of their personality, or their mood in the moment, it is assumed that the number is meaningful: a person with a higher score is more motivated than a person with a lower score. The validity of that claim, just like the validity of any claim from any scientific study, requires evidence. Construct validation is the process by which scientists generate evidence to support a claim that a score produced by an instrument holds a specific meaning (Cronbach & Meehl, 1955). If the numbers used in a study are not valid (e.g., the motivation scores do not reflect levels of motivation) then the conclusions of that study are also not valid.

Modern validity theory and the *Standards for Educational and Psychological Testing* purport that the validation process gives meaning to the interpretation and use of *scores* generated from an instrument, not the instrument itself (AERA, NCME, & APA; 2014; Kane, 2013; Messick, 1989). Thus, validity is not a binary property of an instrument, but instead a judgment made about the score interpretation based on a body of accumulating evidence that should continue to amass whenever the instrument is in use. Accordingly, ongoing validation is necessary because the same instrument can be used in different contexts or for different purposes and evidence that the interpretation of scores generalizes to those new contexts is needed.

There are a host of approaches, specific studies, and corresponding methodologies that can generate construct validity evidence, and we recommend texts that provide thorough treatment on theory and the associated quantitative methods (relevant textbooks; Bandalos, 2018;

Crocker & Algina, 1986; Loevinger, 1957; Markus & Borsboom, 2013; McCoach et al., 2013; Raykov & Marcoulides, 2011; Slaney, 2017). Generally, the process of validation progresses in phases that serve as a means for obtaining different sources of evidence (see Flake et al., 2017 for a concise summary table).

First is the substantive phase, which is used to provide evidence of content and response processes using literature reviews and qualitative item reviews. Examples of studies conducted in the substantive phase are having experts review items and participants think-aloud while they respond to items (Gehlbach & Brinkworth, 2011). Second is the structural phase in which statistical and psychometric properties of those items using quantitative item, factor, and reliability analysis are assessed. Common approaches to obtaining structural evidence are conducting a confirmatory factor analysis to test the intended number and configuration of factors and estimating reliability coefficients for those factors. The final phase is external, in which researchers can test if the instruments' scores behave consistently with theory in relation to other constructs as evidence of relations to other variables. Examples of external phase research are using the multitrait-multimethod matrix (Campbell & Fiske, 1959), and evaluating the predictive power of an instrument for an important outcome. This sequence of validation phases underscores that theoretical *and* statistical evidence supporting the score interpretation is needed *before* using those scores to understand downstream relationships with other constructs or causal relationships (Slaney & Maraun, 2008). Using scores to understand the downstream relationships between constructs is often the goal of a replication study, thus substantive and structural validity evidence are a pre-requisite.

Construct Validity and the Validity of Replication Research

Validity is a broad concept, and the validity of any study conclusion can be evaluated on many aspects, including but not limited to construct validity. In the validity typology described by Shadish et al. (2002), construct validity is a broader concept that includes the validity of experimental manipulations and is one of four fundamental types of validity (the others are internal validity, external validity, and statistical conclusion validity). A strong test of a hypothesis has validity evidence from multiple sources and threats to validity are minimized. While there are various ways to define these different aspects of validity, ensuring that the numbers we use in our data have the meaning we intend, the focus of construct validity as discussed here is a foundational part of the scientific process.

Measurement is foundational to the research process because the ramifications of measurement decisions determine the quality of the collected data, and those data are used to make a conclusion. If a researcher uses scale scores with limited validity evidence, this lack of construct validity cannot be ameliorated through improving statistical practices. Thus, even when replication studies exercise the highest standards of statistical practice, if there is no evidence of construct validity, the results have little meaning. Just as one would not consider the replication of a study with an invalid manipulation meaningful (Fabrigar et al., 2020), one should not consider a replication without construct validity evidence meaningful.

Replication studies have become a mainstay of the reform movement (Plucker & Makel, 2021; Simons, 2014). There exist separate literatures on the practice and merit of replication research and on the theory and quantitative methods for construct validity. What is lacking is an in-depth integration of the two with guidance on how to consider construct validity in the replication process. Until now, the critical role of measurement has received mere mentions (Fabrigar & Wegener, 2016; Flake, 2021; Grice et al., 2017; Loken & Gelman, 2017) and a

recent review of measures used in the Many Labs 2 (ML2; Klein et al., 2014) found that a *majority* of measures lacked validity evidence, especially in the replication sample (Shaw et al., 2020). To empirically evaluate current practices and understand the scope of measurement challenges in replication research, we conducted a systematic review of the measures used in the 100 original and replication studies of the RPP. This is one of the largest measurement reviews of original and replication research based on one of the largest efforts to replicate a set of studies in psychology. We reviewed: (a) validity evidence reported for the measures used in the original studies, (b) validity evidence reported in the replication reports, and (c) measurement challenges encountered by replicators. In the light of our findings, we provide recommendations to integrate construct validity into the replication process and identify future developments needed to advance replication research.

Methods

Procedures and Data

Our sample for this review was every replication report (N = 100) and their original articles (N = 100) in the RPP. We read and coded the validity information for all measures used in the original and replication studies. All RPP materials and our final coding sheet, detailed materials, analysis code and output are hosted on the Open Science Framework (RPP materials: https://osf.io/ezcuj/wiki/home/; our review materials http://bit.ly/for-peer-review).

Coding

For each study in the RPP, we coded the original and replication report to identify the number of measures used, what type of measures they were, and information about the measures that described validity or reliability evidence. We intended to code multiple sources of validity

evidence discussed earlier, such as descriptions of qualitative item pretesting, expert review, or correlational studies between theoretically linked instruments. However, the information available was factor and reliability analyses from the structural phase of validation. Each replication and its respective original article form a pair, with measures and their associated evidence nested within the original and/or replication study.

We employed several methods of quality control to reduce bias and errors made by the coding team (two senior and three junior researchers). We adapted the coding key from Flake et al. (2017), which did not include replication studies, and piloted it on a small sample of articles (n = 11). All remaining articles were individually double coded, followed by a group meeting with a senior researcher to ensure accuracy. Articles that were challenging to code were reviewed and coded by senior coders. Additionally, we used SAS PROC COMPARE to ensure consistency in double-coding.

Results Part 1: Quantitative Overview

Summary of Measures

Figure 1 summarizes the measures used in the RPP and their sources of validity evidence. 362 measures were used in the 100 studies, with a median of two measures per study (SD = 4.47). Item-based scales were the most common type of measure (193, 53% of all measures), of those 96 (50%) were single items and the remainder were multi-item. 40 of the 193 (21%) scales were translated into another language for the replication study. Tasks (i.e.,

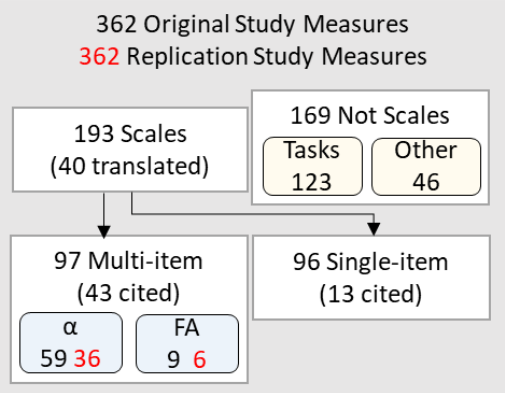

Figure 1. Diagram of measure coding. "Cited" indicates if the scale was reported with a citation, "α" indicates if the scale was reported with Cronbach's internal consistency, and "FA" indicates if the instrument was reported with a factor analysis.

observed behaviors including reaction times or keystrokes) were the second most common measure (123, 34%). The remaining measures (13%) were a diverse group including coded qualitative responses, physiological indicators, coded observational data, and others.

We categorized measures as primary or secondary, details for which are disaggregated in Figure 2. Primary measures (covariates, independent and dependent variables) were used to estimate the targeted replication effect. Secondary measures included manipulation checks, pilot testing, and measures added by the replicators for supplemental or exploratory analyses and these variables were not used in estimation of the replicated effect. Of the 362 measures, 239 (66%) were primary and 123 (34%) were secondary. Of the primary measures, 107 (45%) were item-based scales and 15 of those (14%) were translated into another language for the replication study. 104 (44%) of the measures were tasks, with the remainder falling into a variety of categories. Of the 107 item-based primary scales, 42 were single item scales and the remainder (61%) were multi-item. In the next sections we focus on item-based scales only.

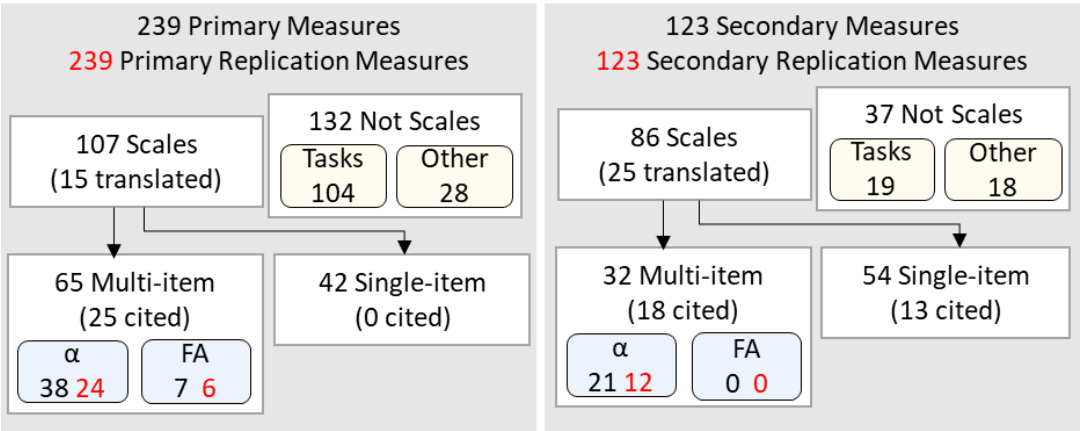

Figure 2. Diagram of measure coding. "Primary" measures indicate that the measure was used to estimate the replication effect, "Secondary" indicates that the measure was collected as a part of the replication study, but not used to estimate the replication effect. "Cited" indicates if the scale was reported with a citation, "α" indicates if the scale was reported with Cronbach's internal consistency, and "FA" indicates if the instrument was reported with a factor analysis.

Within each original and replication study pair, we recorded the number of items, response format, and the scoring method for every scale used in the original and replication

study. A pair was coded to have measurement inconsistency when there was a discrepancy between the original article and the replication report. We recorded 16 of these and detail many of them below. If information about the scale was not in the replication report (e.g., the number of items), we made the liberal assumption that the measure was the same as the original article. Among the 97 multi-item scales across original and replication studies, 83 (85.6%) were reported with the number of items. The average number of items per scale was 9.42, the median was 6, and the mode was 2 (*SD* = 11.34). Most scales had a Likert response format (74%), then a visual analog scale (10%), followed by not reporting the response format (6%).

Summary of Validity Evidence

We coded all available overlapping and unique construct validity evidence in the original study and replication report from the measures and results sections. We only report on three types of evidence (shown in the lower half of Figures 1 and 2). First, scales reported with a citation suggest the presence of previous validity evidence from the scale's development or use. Second, scales reported with a factor analysis suggest having structural evidence. Third, scales with reliability coefficient α suggest evidence of internal consistency. Our review summarizes if the evidence was reported, not the quality of the evidence and the reporting of evidence is not equivalent to the evidence being strong, for example the range of αs for this sample of instruments is .50 to .97. Our results emphasize these sources of evidence over other types of evidence (e.g., item review, test-retest reliability, correlations with other measures) because reporting was so infrequent (< 5 instances, e.g., the correlation between two items was reported for 3 measures) that sparseness precluded them from the final analysis. Critically, though many researchers rely on Cronbach's α reliability coefficient as a sole source of evidence to support a score, a reliability coefficient does not signal validity (see Schmitt, 1996 for a thorough

discussion).

Original Report

The RPP used 193 scales in total, and 56 of those were reported with a citation in the original study (29%). It was rare for single item measures to be reported with a citation (13 out of 96 single item measures; 14%), whereas 43 of 97 (44%) of multi-item scales accompanied a citation, suggesting that these measures were previously developed. Of the 97 multi-item scales, 59 were reported with a reliability coefficient and 9 with a factor analysis (see bottom portion of Figure 1). Collapsing across any source of evidence, Figure 3 shows that 20% of the 97 scales were reported with no evidence, whereas 80% were reported with at least one source of evidence.

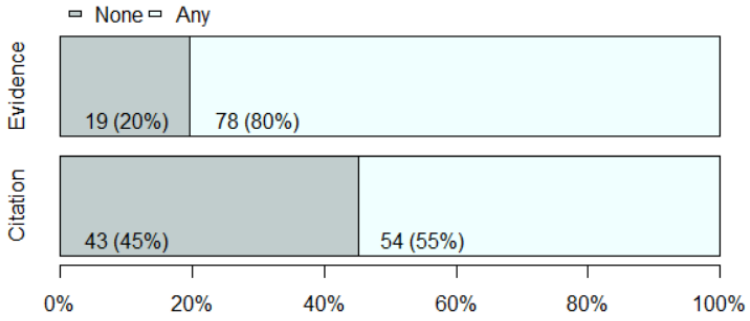

Figure 3. Frequencies and proportions of scales with more than 1 item in terms of citations and any validity evidence in original studies.

Across all these scales, including single-item ones, 107 were primary, 65 (61%) of those were multi-item scales and 25 (39%) were reported with a citation. None of the single item primary scales were reported with a citation. Figure 4 presents the frequency of α reliability coefficients and a factor analysis for all multi-item scales (N = 97), disaggregated by if they were included in the estimation of the replicated effect (primary) or collected during the replication study but for some other reason (secondary).

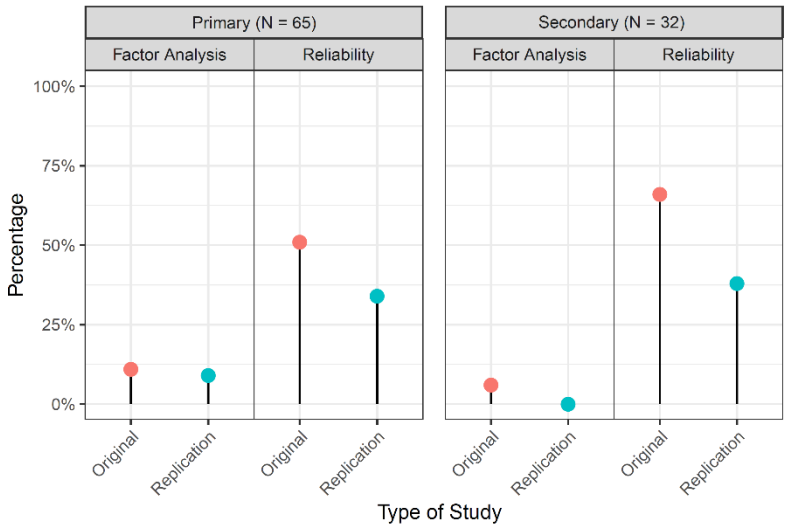

Figure 4. Lollipop graph showing the percentage of reliability coefficients and factor analyses reported for primary and secondary scales in original and replication studies.

Replication Report

Of the 97 multi-item scales, 37 (38%) of those were reported with additional validity evidence in the form of psychometric analyses in the replication report. Specifically, replicators reported a reliability coefficient or factor analysis estimated from the replication sample. However, of the 54 scales that had no citation in the original, 29 were reported with no other validity evidence in the replication report. Thus, slightly more than half of the scales began the replication process with little or no validity evidence in the original publication *and* no validity evidence from the replication sample. Of the primary measures more specifically, 20 had no citation in the original and no additional information in the replication report. Note that a single factor model with equal loadings is necessary for the valid interpretation of α reliability coefficients, which are the most common psychometric information reported.

Of the 40 translated scales (including 1-item measures) used in the replications, 8 had a citation to an existing validity study that described evidence for the translated version, whereas the rest (80%) were translated by the study authors. Fifteen translated scales were used as primary measures, and only one of those had a previously cited translated version. In sum, measures used in replication studies had less reported validity evidence than the original studies.

Results Part 2: Qualitative Summary with Examples from the RPP

From the results above, we identified four broad measurement challenges that replicators are likely to face during the replication process: (a) limited information about measures, (b) limited validity evidence for scales, (c) measurement differences from the original to the replication study, and (d) translation of instruments into other languages. In the next section, we illustrate these challenges with examples from the RPP. These specific examples elucidate how

measurement challenges are realized in the theoretical, methodological, and practical aspects of conducting and interpreting replication research. After reporting on these challenges, we turn to discussing why they need to be addressed and how to address them.

Challenge 1: Original study provides limited information about measures

Upon reviewing a study's methodology to plan a replication, researchers are likely to find that key information about the study's measures is missing or difficult to discern. In our experience, ascertaining how the variables were measured was a challenging and sometimes impossible task because construct definitions and how they were measured by scales were not clearly or completely reported. Of the scales used in the RPP (n = 193), a nontrivial percentage lacked key information: number of items not reported (8%), response format not reported (6%), and how the scale was scored not reported (scales with more than 1-item, 51%). Replicators explicitly described this lack of information. In study 46, the wording of the items was unclear, and replicators later found that they were incorrectly reported in the original study (Original Article: Exline et al., 2008). In another report, replicators described being unable to ensure the same scoring as the original (RPP Study 91; Original Article: Moeller et al., 2008). These problems occurred after due diligence was taken to gain access to the original study materials and even when original study authors cooperated.

Challenge 2: Primary measures have no or little validity evidence

Of all the scales used in the reproducibility project, 71% had no citation to published validity evidence in the original study. Further, for the primary measures with more than one item, 20% were used without any information; that is, these longer measures had no citation, or any other validity or reliability evidence reported from data collected in the original study. Thus,

many scales appear to have been created on-the-fly, lacking multiple sources of construct validity evidence. This is consistent with other reviews of measurement practices in psychology (Flake et al., 2017; Slaney, 2017). Thus, researchers often encounter a scenario where they set out to replicate a study and the primary variables of interest are measured with scales that have no or little existing evidence.

Challenge 3: Measurement Differences

Even when strong construct validity evidence is reported in an original study, measurement challenges can occur. Our review indicates the presence of measurement inconsistency between original and replication studies: in 16 of the 100 there was at least one measurement difference reported. These measurement differences manifested in a variety of ways with some stemming from instrument modification whereas other differences, for which we provide examples of below, became apparent in the data analysis process even though the instrument was not modified. Inadvertent scale modifications were caused by Challenge 1 (limited information about scales). Changes to the item wording, presentation, and response options introduces measurement differences that may interfere with the comparison of the scores from original to replication study. Less obvious is the presence of measurement differences when replicators administered the same scale and followed the same experimental protocol as the original study. In such instances, the construct validity evidence (usually in the form of a psychometric analysis) indicates that the replication study may not have been measuring the same construct as the original study.

Reliability Differences

Study 92 was a replication of Murray et al. (2008), which concluded that the desire for

connectedness in a romantic relationship magnifies rejection concerns. This study included six multi-item scales, and for five of those, the reliability estimate in the replication sample was lower than the original. The replicators conducted a reliability analysis and removed items to achieve a higher α in the replication sample. In one instance, the replicators noted that even with item removal, the α was substantially lower than in the original (original α = .90, replication α = .77). In addition to differences in reliability from original to replication, item removal could decrease the content validity of the instrument, causing the replication study to have less evidence of content than the original. Based on the targeted statistical test for the replicated effect, findings were interpreted as mixed.

Item Correlation Differences

Reinhard (Study 41 in RPP) conducted a replication study of Förster et al. (2008), which concluded that priming people with aggression can assimilate aggression into their later social judgements and this priming effect varies with cognitive processing styles. The primary dependent measure in this study was how aggressive a target was rated, based on the average of responses from two items. In the original paper, the two items were strongly correlated (r = .63, p not reported), but these items were not significantly correlated in the replication (r = .17, p = .16). A large correlation justifies the averaging of responses of the two items because their shared variation could be attributed to a common construct, whereas a weak correlation introduces uncertainty as to whether the items are measuring the same thing. Confronted with this conflicting information, the replication author followed the original study protocol and averaged the two items to measure the dependent variable, but also conducted a follow up analysis of each item separately. Based on the targeted statistical test for the replicated effect results were interpreted as a failure to replicate.

Different Factor Solutions

Study 7 in RPP was a replication of Monin et al. (2008), where it was concluded that those who take an unpopular, but morally laudable stand on an issue ("a moral rebel") would cause others to feel threatened when others' behavior was in contrast with that of the 'moral rebel.' This study included 13 scales, three of which were self-reported negative affect, positive affect-high arousal, and positive affect-low arousal. In the original study, these subscales were derived from a factor analysis. However, when the replicating author conducted an exploratory factor analysis, they observed two instead of three factors: one factor included all positive affect items and the other negative affect items. The replication author followed the original study protocol and aggregated the items based on the original study's factor analysis. Note that scores representing positive affect-high arousal and positive affect-low arousal in the replication study are not supported by factor analysis results. There is a tension between using the scales in the same way as the original versus using the scales supported by the factor analysis results, making it unclear how to proceed. Based on the targeted statistical test for the replicated effect, results were interpreted as a failure to replicate.

Challenge 4: Translation

The RPP was an ambitious replication effort, with participating laboratories from different regions of the world. Following this trend, even larger-scale international projects are underway or have been completed (Klein et al., 2014; Moshontz et al., 2018). These efforts directly address the shortcomings of small, non-representative samples, but like the RPP, they present complex measurement challenges. A common challenge faced by replicators in the RPP was that the studies required translating measures into different languages for the first time. We found that 40 of the scales used in the RPP were translated. Only 8 (20%) of these translations

adopted a previously translated version with validity evidence.

Scale translation is an onerous process that poses unique challenges to construct validity. When measures were translated from the original for the replication study, the replicators in the RPP were thorough, usually utilizing multiple translators and back translation techniques. However, scores from translated scales require validation after the initial translation to confirm that translated items are capturing the construct in a comparable way. For example, validity studies on translated versions of the Positive and Negative Affect Schedule (PANAS; used in 2 of studies in the RPP) have found a lack of measurement equivalence (Lim et al., 2010; Terracciano et al., 2003), indicating that scores from the original version are not comparable to scores from the translated version.

Discussion

The RPP remains one of the largest replication efforts ever undertaken by psychological scientists, which raised concerns about the replicability of results. The RPP ignited a methodological reform movement that has spurred the improvement of research practices and the open data and materials have enabled the review we report here. Our review of the 100 original and replication studies from the RPP demonstrates that construct validation, foundational in the scientific process, is severely neglected in original and replication research, which makes replication research difficult to conduct and compromises validity. In this discussion, we summarize the ramifications of these practices and offer three recommendations for improvement that will bolster the validity of original and replication research. Then we discuss future methodological research that will improve the conduct of the next generation of replication studies and these studies' value to the scientific community and society.

Current Practices

Our review is consistent with previous meta-scientific research: psychological scientists use scales with little or no reported validity evidence (Slaney, 2017). In the RPP, 44% of scales used to measure variables that were included in the analysis of the replicated effect had no supporting citation to a construct validation study in the original work. Then the replication studies, using the same measures, inherit this very lack of evidence. It is not common practice to evaluate construct validity in replication studies, which results in lower validity evidence than the original works (see Figure 4).

Our emphasis on construct validity might lead one to ask, "if construct validity is improved, will replicability improve?" The answer is complex because replication rates are multiply determined. Strong construct validity evidence, just like other sources of validity evidence, strengthens the test of a hypothesis. But validity evidence cannot determine the outcome. Stated differently, construct validity is a necessary but insufficient condition of replication. If an original study possesses strong validity, it provides the means for a researcher to observe a meaningful effect, if it exists. True effects, detected by studies with strong validity evidence, should replicate. However, in practice, researchers do not know if original studies reported true effects (i.e., they could be Type I errors). Thus, when replicating studies that reported Type I errors, even if those studies have valid scores (construct validity), ideal experimental manipulations (internal validity), representative samples (external validity), and aptly applied statistical procedures (statistical conclusion validity), they should not replicate. Those studies will, however, provide more conclusive evidence of the (null) effect.

Conversely, some studies with fatal internal validity or construct validity flaws do replicate (e.g., Chen & Risen, 2010; see Yarkoni, 2021 for a discussion of invalid yet replicable

results). The question we raise in this review is, “if there is no construct validity, what can a replication study tell us?” And the answer is: little. Replicability tells us how consistent a finding is; it does not tell us if the finding is valid. Validity tells us how meaningful a finding is; it does not tell us if the finding will replicate. Replication studies, like all studies, should emphasize producing a conclusion with limited threats to validity (internal, external, statistical conclusion, and construct; Shadish et al., 2002). If psychological scientists improve on all forms of validity in their research, including construct validity, original studies will be better tools for detecting true, meaningful, and replicable effects. Replications of such studies will be stronger tests of theory, which will increase their contribution to the accumulation of knowledge. In the following sections we recommend ways that the community can work toward improving construct validity.

Recommendation 1: Improve the Validity and Transparency of Original Research

Our review indicates that a central challenge of conducting replication research is a lack of transparency and validity of original research. Original studies that lacked essential information made direct replication studies difficult to conduct. There were multiple occasions when replicators could not locate and administer the same materials and protocol. Further, original research also lacked strong evidence of construct validity, with many researchers relying on the α reliability coefficient as a sole source of evidence to support a score. Reliability coefficients are not evidence of validity: scores can be reliable (i.e., consistent) but not convey the meaning intended (Schmitt, 1996). The research community needs to continue to improve the conduct and reporting of construct validity in original research (see Flake & Fried, 2020 for a detailed discussion). Detailed information and strong evidence for instruments is not only required to facilitate replication, but also to evaluate the validity of the original research results.

Recommendation 2: Consider Construct Validation in the Selection, Design, and Interpretation of Replication Research

Before the transparent reporting of strong construct validity evidence in original research becomes commonplace, scientists will still encounter original work that uses scores with little or no validity evidence that they wish to conduct a replication of. In fact, this review and others (Slaney, 2017) suggest that replicators are *likely* to be in this very situation. While replications can be powerful tools for demonstrating the reliability of original effects, some original effects rest on such shoddy foundations that spending resources to replicate them is futile (Flake et al., 2021; Yarkoni, 2021). A concrete take-away is that scientists should only select studies for direct replication that provide a severe test of the hypothesis (Mayo, 2019) via a thorough review of the validity evidence (construct, statistical, internal, and external). Concretely, if an original study reports limited substantive or structural construct validity evidence, those threats to validity will transfer to the replication study. Large-scale replication studies such as the RPP and Many Labs have focused on selecting studies from certain journals or time periods based on novelty or impact. While this may have been effective for garnering attention, a balanced approach that justifies selection based on the strength of the original evidence alongside those other considerations would result in replication studies that are a good return on investment by generating results with fewer threats to validity.

If researchers determine there is very limited validity evidence reported in the original study and still choose to replicate the study, we recommend that replicators gather evidence of construct validity. There are many potential sources of evidence and what constitutes strong construct validity evidence for scores from an instrument will depend on how the instrument is being used. At minimum, replicators should incorporate basic checks of validity and reliability into a replication study to rule out serious threats to the conclusion. Examples include a review

of items or stimuli to make an evaluation of face and content validity, checks on assumed scoring structure when applicable (e.g., using a component or factor analysis), and reporting of an appropriate measure of reliability. These checks are akin to conducting a manipulation check or developing a sample size plan and will go some distance toward increasing the impact and validity of the replication study. Some construct validity evidence is more than none, and transparent reporting of the strength of the evidence in a replication study is an important first step toward strengthening the validity of replication research.

Another option is to improve upon the measures (e.g., use different or redeveloped instruments), which would preclude a direct replication. Results from such a study would serve as a stronger and different test of the original claim, but still evaluate the generalizability of the phenomenon (Shrout & Rodgers, 2018). While there is value in direct replication studies, there is limited value in running a study that will return results open to alternative explanations linked to poor construct validity. Thus, partial or approximate replication research with less emphasis on statistical significance can build upon the evidence provided by the original work (Anderson & Maxwell, 2016).

Regardless of whether the replication is direct or approximate, all validity evidence should be considered when interpreting the result. If construct validity evidence is weak, like other threats to validity (e.g., confounds), that should be explicitly acknowledged in the final interpretation of the replication effect. To the extent that replications generate new knowledge about the constructs or downstream effects on them, theory has been fortified, updated, or extended. Stated differently, such knowledge adds to the generalizability of the constructs under study and how it does or does not provide evidence for the original effect (Flake et al., 2021; Shrout & Rodgers, 2018). In discussions about failed replications in psychology, theory failure,

hidden moderators, and questionable research practices are often debated (Finnigan & Corker, 2016; LeBel et al., 2017) with little attention to the foundational issue of measurement.

Recommendation 3: Plan for Measurement Differences

A successful implementation of a replication study assumes that constructs are the same across original and replication studies and the measures between studies are equivalent (Fabrigar & Wegener, 2016). Imagine running an experiment testing the effect of a diet on weight loss. Participants were randomized into an experimental diet or a control diet, and on average the participants on the experimental diet lost more weight. Next, a replication of this experiment in another sample drawn from the same population is conducted. However, in this second experiment, a low battery in the scale causes intermittent, erroneous readings of participants' weight and no difference in the mean weight between the diet groups is detected. It is difficult to interpret the effectiveness of the diet in the replication study because the scale is functioning differently (i.e., is non-equivalent) from the original study. In fact, readings from the scale are invalid, causing the replication study conclusions to be invalid. Measurement equivalence is necessary for results from the original and replication studies to be comparable.

Scores from a measure are equivalent when their properties do not differ across time, groups, or settings (Meredith, 1993) and it is this property of measurement equivalence that enables comparability of effects across studies. However, replication studies take place later in time, often in more diverse geographic locations, and increasingly, in another language; all of which can contribute to measurement non-equivalence. Broadly, measurement non-equivalence occurs when participants interpret scales differently and this results in different psychometric properties, such as when items or stimuli correlate differently (forming different components or factors) or exhibit different levels of reliability. Even when replicators in the RPP made efforts to

homogenize the settings in study implementation or rigorously translate a measure into another language, they still observed measurement differences.

Replicators should expect translated instruments to exhibit measurement non-equivalence. This challenge is well documented in international achievement testing and studies often report a substantial number of non-equivalent items (e.g., 18-54%, Gierl & Khaliq, 2001), also referred to as differentially functioning items. Translated versions of the PANAS, used in multiple studies in the RPP, have demonstrated non-equivalence (Lim et al., 2010; Terracciano et al., 2003). This complication is compounded when replication studies are conducted in multiple languages, which is becoming a norm. In the Many Labs 2 project, replicators translated instruments into 16 languages, which were used across dozens of labs and the reliability of the translated versions was lower on average when compared to instruments administered in the language in which they were developed (Shaw et al., 2020).

Evaluate Measurement Equivalence

There is an abundance of research demonstrating how to statistically evaluate measurement equivalence and instrument translation in the testing and psychometrics literature that can be integrated into the replication process (Allalouf et al., 1999; Meredith, 1993; Solano-flores et al., 2009; van de Schoot et al., 2012). Examples include using item response theory or confirmatory factor analysis to identify differential item functioning, in addition to think a-loud protocols to ensure participants are understanding items in the intended way or to confirm accurate translation. With evidence of measures' equivalence across different languages these replication projects can produce large, public datasets that promote valid instrument use across diverse populations and support increasing the generalizability of conclusions.

Though replication studies have not formally evaluated measurement equivalence from original to replication study or across translated versions, our review and a review of the ML2 instruments (Shaw et al., 2020) suggests the presence of non-equivalence. When evidence points to a failure to achieve comparable foundational aspects between the original and replication study, it is not sensible to interpret the downstream test of the hypothesis from the replication study. For example, if the manipulation in a replication of an experiment failed, one would not progress to conduct a statistical test between the treatment and control groups. Observing measurement differences between original and replication studies affords an explanation to replication failure beyond Type I errors. These replication failures can point to future directions of research that can systematically examine the generalizability of constructs and their effects, potentially uncovering theoretically important heterogeneity.

Next Steps for Replication Research

The future of replication research will involve studies even bigger than the RPP and will require methodological advancement. To improve the next generation of replication research, psychological scientists can focus on expanding infrastructure by working to accumulate results systematically and developing measurement methodologies and practices for large-scale replication.

Large-scale Replications

First, even a single, well designed replication study is limited in what it can tell us. Just like original studies can be flawed or produce false positive results, so too can replication studies. Practices are shifting to the next generation of replication research to take a team science approach. Initiatives like the Many Labs and the Psychological Science Accelerator (Moshontz

et al., 2018) provide a way for large groups of researchers to work together to collect big and more representative datasets. These data can be used to thoroughly evaluate the generalizability of constructs and the downstream effects connected to them. However, advances will be needed in developing the infrastructure to complete this work and the system for how to incentivize, fund, and credit it (Coles et al., 2022). Another avenue is to build from the progress in sharing data and materials psychologists have made so far to develop platforms for contributing site-specific replication data that could later be pooled for larger-scale and more representative evidence for constructs and effects.

Measurement Equivalence

Measurement equivalence has a long history of theoretical and technical developments in the field of psychometrics, which can offer a starting point for quantifying measurement differences in replication research. The assumption of measurement equivalence can be empirically assessed with scales, questionnaires, and tests using various quantitative methodologies (see Millsap, 2011 for an overview), where psychometric properties are tested for equivalence across a series of successively more restrictive models. There are multiple approaches for scoring instruments to take into account measurement non-equivalence in downstream analyses (Devlieger et al., 2016). These methodologies could be applied in replication research. Further, there have been significant advances in modeling for large and complex data structures, as well as the compilation of data from many studies (Curran, 2009). Multilevel structural equation modeling and item response theory (e.g., Muthén & Asparouhov, 2013) can accommodate item level data from instruments that are clustered by multi-lab replication studies and facilitate measurement equivalence testing. Advances in integrative data analysis and meta-analysis have motivated the development of commensurate measures for cross

study compilation (e.g., Curran et al., 2009). We have yet to see these methods applied and adapted for the replication context. Further methodological research can evaluate which of these methods can be readily adopted to enhance the comparability of results from replication studies. These existing methodologies, paired with larger datasets, have the potential to refocus replication research on understanding the generalizability of the constructs under study and the effects associated with them.

An omission in the literature on improving the replicability of results is how to operationalize the equivalence of measurement properties between original and replication studies. The field cannot continue to ignore measurement non-equivalence as it moves toward larger-scale replication research that requires the translation. Measurement equivalence is not a binary property of an instrument (Horn & McArdle, 1992) and some items can be equivalent while others are not (Byrne et al., 1989). Previous work on how a lack of measurement equivalence influences downstream prediction (Millsap, 1995) and contributes to differential validity (Drasgow, 1982) can be extended to the purpose of planning and interpreting results from replication studies when measurement differences arise. While there are still no clear guidelines for replicators to follow, we hope that our review builds interest in developing theory and practice on measurement equivalence for replication studies.

Limitations

First, the validity evidence presented is a convenience sample based on the scales used in the RPP. Although our findings are consistent with measurement reviews in other areas of psychology, these findings may not generalize to a larger population of studies that will undergo replication. Further, coding was an interpretive process, and we encourage others to examine and re-examine our open materials to further enrich our collective understanding—or rectify our

collective misunderstandings.

Conclusion

Construct validation is at the foundation of valid and replicable psychological science. Our review of the construct validity evidence from the 100 studies in the RPP indicates that construct validity evidence is lacking in original and replication research, limiting the impact of both. Replicability without validity will not do enough to improve the state of psychological science. We argue for closer inspection of construct validity in the selection and design of replication studies and their corresponding original studies, highlighting that validity evidence must first generalize from original to replication studies before meaningful replication can be achieved.

Psychology is growing from a critical self-evaluation of our claims and underlying methodologies. As a scientific discipline, we are taking steps to becoming more transparent, more thoughtful about our statistical practices, and more rigorous in the way we conduct science and report our results. Such methodological revolutions (Rodgers & Shrout, 2018) or periods of renaissance (Nelson et al., 2018) are indications that the discipline of psychology is growing. The purpose of the current work is to extend the ongoing evaluation of our methodology by highlighting measurement practices that present real challenges when replicating research. Due to the current lack of transparency and limited application of rigorous methodologies for measurement, results from replications continue to be limited in their conclusions. To address these challenges, we need to rediscover existing methodologies for psychological measurement, implement them in our original and replication studies, and develop methodologies for the next generation of replication research.

References

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of translation DIF on verbal items. *Journal of Educational Measurement*, *36*(3), 185–198. https://doi.org/10.1111/j.1745-3984.1999.tb00553.x

American Educational Research Association, National Council on Measurement in Education, & American Psychological Association. (2014). *Standards for educational and psychological testing*. Joint Committee on Standards for Educational and Psychological Testing.

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1–12. https://doi.org/10.1037/met0000051

Bandalos, D. L. (2018). *Measurement theory and application for the social sciences*. Guildford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures : The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Chen, M. K., & Risen, J. L. (2010). How choice affects and reflects preference: Revisiting the free-choice paradigm. *Journal of Personality and Social Psychology*, *99*(4), 573–594.

Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H., & Altschul, D. (2022). Build up big-team science. *Nature*, *601*, 505–507.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winstron.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological science. *Psychological Bulletin*, *52*, 281–302.

Curran, P. J. (2009). The seemingly quixotic pursuit of a cumulative psychological science: Introduction to the special issue. *Psychological Methods*, *14*(2), 77–80. https://doi.org/10.1037/a0015972

Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., Sher, K., & Zucker, R. (2009). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Applications Series*, *49*(3), 214–231. https://doi.org/10.1080/00273171.2014.889594

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods. *Educational and Psychological Measurement*, *76*(5), 741–770. https://doi.org/10.1177/0013164415607618

Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin*, *92*(2), 526–531. https://doi.org/10.1037/0033-2909.92.2.526

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*(4),

316–344. https://doi.org/10.1177/1088868320931366

Finnigan, K. M., & Corker, K. S. (2016). Do performance avoidance goals moderate the effect of different types of stereotype threat on women ' s math performance ? *Journal of Research in Personality*, *63*, 36–43. https://doi.org/10.1016/j.jrp.2016.05.009

Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist*. https://doi.org/10.1080/00461520.2021.1898962

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 1–22. https://doi.org/https://doi.org/10.31234/osf.io/hs7wm

Flake, J. K., Luong, R., & Shaw, M. (2021). Addressing a crisis of generalizability with large-scale construct validation. *Behavioral and Brain Sciences*.

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), 1–24. 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. https://doi.org/10.1037/a0025704

Gierl, M. J. ., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests : A confirmatory analysis. *Journal of Educational Measurement*, *38*(2), 164–187.

Grice, J., Barrett, P., Cota, L., Felix, C., Taylor, Z., Garner, S., Medellin, E., & Vest, A. (2017). Four Bad Habits of Modern Psychologists. *Behavioral Sciences*, *7*(3), 53. https://doi.org/10.3390/bs7030053

Horn, L. J., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144.

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., … Coles, N. A. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*. https://doi.org/10.1038/s41562-020-01007-2

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., 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*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

LeBel, E. P., Campbell, L., & Loving, T. J. (2017). Benefits of open and high-powered research outweigh costs. *Journal of Personality and Social Psychology*, *113*(2), 230–243. https://doi.org/10.1037/pspi0000049

Lim, Y.-J., Yu, B.-H., Kim, D.-K., & Kim, J.-H. (2010). The Positive and Negative Affect Schedule: Psychometric properties of the Korean version. *Psychiatry Investigation*, *7*(3),

163. https://doi.org/10.4306/pi.2010.7.3.163

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

Mayo, D. G. (2019). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

McCoach, D. B., Gable, R. K., & Madura, P. J. (2013). *Instrument development in the affective domain: School and corporate applications*. Springer.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invarinace. *Psychometrika, 58*(4), 525–543.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30*(4), 577–605. https://doi.org/10.1207/s15327906mbr3004_6

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Moshontz, H., Campbell, L., Ebersole, C. R., Ijzerman, H., Urry, H. L., Forscher, P. S., Mccarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake,

J. K., & Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*.

Muthén, B. O., & Asparouhov, T. (2013). *New methods for the study of measurement invariance with many groups*. statmodel.com

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*, 511–534.

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

Plucker, A. J., & Makel, M. C. (2021). Replication is important for educational psychology: Recent developments and key issues. *Educational Psychologist*, 1–35.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Rodgers, J. L., & Shrout, P. E. (2018). Psychology's replication crisis as scientific opportunity: A précis for policy makers. *Policy Insights from the Behavioral and Brain Sciences*, *5*(1), 134–141. https://doi.org/10.1177/2372732217749254

Schmitt, N. (1996). Uses and abuses of Cronbach's alpha. *Psychological Assessment*. https://doi.org/10.1037/a0040195

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Wadsworth Publishing.

Shaw, M., Cloos, L. J. R., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications : Insights from Many Labs 2. *Canadian Psychology*, *61*(4), 289–

298. https://doi.org/10.1037/cap0000220

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 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*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

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

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at Perspectives on Psychological Science. *Perspectives on Psychological Science*, *9*(5), 552–555. https://doi.org/10.1177/1745691614543974

Slaney, K. L. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions* (J. Martin (ed.)). Palgrave Macmillian. https://doi.org/10.1057/978-1-137-38523-9

Slaney, K. L., & Maraun, M. D. (2008). A proposed framework for conducting data-based test analysis. *Psychological Methods*, *13*(4), 376–390. https://doi.org/10.1037/a0014269

Solano-flores, G., Backhoff, E., Contreras-niño, L. Á., Backhoff, E., Contreras-, L. Á., & Solano-flores, G. (2009). *Theory of Test Translation Error Theory of Test Translation Error*. *5058*(May). https://doi.org/10.1080/15305050902880835

Terracciano, A., McCrae, R. R., & Costa Jr., P. T. (2003). Factorial and construct validity of the

Italian Positive and Negative Affect Schedule (PANAS). *European Journal of Psychological Assessment*, *19*(2), 131–141. https://doi.org/10.1027//1015-5759.19.2.131.Factorial

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Yarkoni, T. (2021). The generalizability crisis. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X20001685