2020-06-10

Reliability and Validity Practices in Randomized Controlled Trials: Current Trends and Recommendations

Jennifer A. Z. Romano
Brigham Young University

Reliability and Validity Practices in Randomized Controlled Trials: Current Trends and Recommendations

Jennifer A. Z. Romano

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Scott Baldwin, Chair
Michael Larson
Scott Braithwaite

Department of Psychology

Brigham Young University

ABSTRACT

Reliability and Validity Practices in Randomized Controlled Trials:
Current Trends and Recommendations

Jennifer A. Z. Romano
Department of Psychology, BYU
Master of Science

The verity of conclusions drawn from psychological research hinges on the reliability and validity of the measures used to collect the data. Any research conducted using measures with low reliability or validity is rendered essentially useless; thus, reporting reliability and validity evidence for measures employed in research is an essential component in creating rigorous, replicable research. Multiple reporting standards have been implemented and revised over the years with the intent to improve measurement and reporting practices within clinical psychology, though few guidelines have been suggested regarding adequate reporting practices for studies' measures. We reviewed a representative sample of randomized clinical trials (RCTs) published in the *Journal of Clinical and Counseling Psychology* in 1994, 2002, 2010, and 2018 for reported reliability and validity evidence. We examined whether the implementation of reporting standards led to improvement in reporting measures' reliability and validity evidence over time, along with how frequently articles recently published in one of the top clinical psychology journals reported reliability and validity evidence. We found that only 58.1% of measures used in articles published in 2018 reported reliability evidence, and only 12.4% reported validity evidence. Furthermore, although reporting of reliability and validity evidence has improved when comparing articles published in 2018 to those published in 1994 or 2002, such reporting practices were not significantly different from articles published in 2010. We provide a discussion of the importance of these findings and recommendations for improving reporting practices in future research.

Keywords: measurement, reliability, validity, reporting practices

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Reliability and Validity Practices in Randomized Controlled Trials: Current Trends and Recommendations

Psychology is in the midst of a replication crisis (Lilienfeld, 2017; Pashler & Wagenmakers, 2012; Shrout & Rodgers, 2018; Spellman, 2015). Psychological scientists and practitioners alike rely on published studies to guide best treatment practices and to identify evidence-based treatments; when study results do not replicate, their findings are thrown into question. A reproducibility study of 100 studies demonstrated that the current body of psychological research has low replicability: only 36% of replications had significant results, while 97% of the original studies had significant findings. In addition, the 95% confidence interval of the replication effect size only contained 47% of the originally reported effect sizes (Open Science Collaboration, 2015), indicating that reported effect sizes replicated in fewer than half of the studies. Though the replication crisis began with the frequent failure to replicate studies published in social and cognitive psychology, subsequent studies also show replication failures in clinical psychology (Munafò et al., 2017; Tackett et al., 2017).

The replication crisis has led to individuals proposing many different solutions. In an effort to improve the reproducibility of findings and the transparent reporting of study designs, many have begun to call for researchers to preregister their studies (Munafò et al., 2017; Nosek et al., 2018), an often time-consuming process with few researcher incentives (Murray et al., 2019). The preregistration of studies is being adopted at an accelerated rate in psychology (Nosek & Lindsay, 2018). Even with incentives changing to reward preregistration (Nosek & Lindsay, 2018), many studies are still being published without preregistration. For example, Murray et al. (2019) found that despite increased preregistration incentives and the risk of being fined for failing to report transparent summary data, less than 10% of anorexia nervosa treatment articles published in peer-reviewed journals demonstrated evidence of preregistration. Others have advocated for increased training in statistics, measurement, and methodology; Aiken, West, and Millsap (2008) indicated that most Ph.D. programs included in their study only required their students to take one introductory level statistics course. With such little training in statistics and measurement, the lack of reproducibility in clinical psychology research may be contributed to, at least in part, by researchers not fully attending to measurement.

Measurement in psychology and other social sciences has many challenges. In psychology, researchers often study constructs which cannot be directly measured. They must first create an operational definition of the construct, which may not capture every facet of the construct; or, if it successfully captures every facet, it may require such a heterogeneous and complex measure that the time it would take to employ the measure may outweigh its theoretical utility. In addition, many psychological constructs are related, even overlapping with one another, and creating meaningful distinctions between them is difficult (Naragon-Gainey et al., 2018). Different measurement methods, such as the frequently employed self-report method, also have limitations that may prevent researchers from fully understanding the construct of interest (Wood et al., 2001). For example, self-report measures rely on an individual’s level of insight and may often be influenced by response biases and social desirability (Nisbett & Wilson, 1977). In addition, validity evidence for a measure is difficult to find and is often generated by comparing a new measure with an older measure of the same construct, rather than an external factor that could validate the measure (Clark & Watson, 2019).

Poor measurement practices, including the failure to report and replicate validity and reliability evidence, can influence research in various ways. Estimates of reliability and validity are dependent on the population and context being studied (Clayson & Miller, 2017; Smith &

McCarthy, 1995); therefore, an estimate obtained from one population or context may not replicate in another population or context. Clayson & Miller (2017) give the example that the reliability and validity scores of a measure of depressive symptoms in an undergraduate sample may not generalize to other populations such as a geriatric population or other contexts such as an outpatient clinic. Even in similar samples (e.g., an undergraduate population in two different universities), other contextual factors may differ between the samples in such a way that the reliability and validity estimates obtained in one sample fail to generalize across samples. If the scores from a measure have poor validity, our conclusions from the research we conduct may be erroneous, as we may not be measuring what we purport to measure. We may never know this if we do not repeatedly calculate and report the validity of the measure each time it is used. If a scores from a measure demonstrate high reliability in one sample but poor reliability in other samples, studies using this measure will not have consistent results. Loken and Gelman (2017) addressed the common misconception that measurement error always reduces effect sizes, demonstrating that studies with small sample sizes (especially those with fewer than 500 participants) frequently have inflated effect sizes due to measurement error. Such measurement error often arises when measures have low reliability; thus, the lack of reporting reliability evidence may result in false-positive results that fail to replicate in future studies.

Just as researchers have many possible decisions to make that may lead them to analyze data differently while attempting to answer the same research question, as illustrated in a study conducted by Silberzahn et al. (2018), researchers also have many possible decisions to make regarding measurement. For example, many of the measures utilized in depression research use a design in which symptoms are summed to generate a depression score. However, Fried and Nesse (2015) identified 1030 unique symptom profiles of outpatients with depression. With such heterogeneity in the symptom presentation of depression, simply generating sum-scores and classifying depressed individuals as those who meet a certain threshold may not represent the full construct of depression, thus threatening the construct validity of the depression measure. Such a measure may produce false-negative findings in studies, simply because it fails to capture the true heterogeneity of depression presentation. In a similar manner, measures normed on a sample with a homogenous symptom presentation of depression will fail to have acceptable construct validity by failing to capture the true heterogeneity of depression presentation, and its findings may not replicate across samples of individuals with a depression symptom presentation that is different from the original sample.

Flake and Fried (2019) identified many questionable measurement practices (QMPs) that might threaten the validity of a study's conclusions. They explained that QMPs prevent researchers from being able to identify such threats. Common QMPs include the creation of measures that have never been used before, the failure to report reliability and validity evidence for the measures employed, and the omission of analyzed scales from published research. They provide specific questions designed to promote transparency and improve the rigor of measurement practices, including questions about why measures were selected, why and how measures were modified, and whether the measure was created "on the fly" (p. 9) along with justification for creating the new measure.

Similar to Flake and Fried's (2019) emphasis on avoiding QMPs, one of the most common responses to the replication crisis has been an increased emphasis on transparent reporting practices. Lack of transparency at any stage of a study, including study design, data collection, measurement, and analysis, often culminates in researchers engaging in questionable research practices (John et al., 2012; Simmons et al., 2011) or, in more extreme cases, p-hacking

(Vazire, 2017). In a survey that incentivized truth-telling, John et al. (2012) reported that 94% of researchers admitted having committed at least one QRP. Regardless of whether the practice is judged to be justifiable by the researcher, transparent reporting of such practices is key. Failing to disclose these practices in the past threatens to erode trust in science, as consumers of scientific research find themselves unable to distinguish between psychologically rigorous research and its more questionable counterpart (Vazire, 2017).

Efforts to improve the transparency of clinical trials have been ongoing since the 1990s. Psychologists have implemented guidelines such as the *Consolidated Standards of Reporting Trials* (CONSORT; Schulz, Altman, & Moher, 2010) and the *Journal Article Reporting Standards* (JARS; Appelbaum et al., 2018) to help increase the rigor of clinical trials and improve the reporting of psychological research generally. The most updated CONSORT Statement (Schulz et al., 2010) was designed as a guideline for reporting practices of randomized clinical trials (RCTs). It includes a 25-item checklist of information (p. 699) for researchers to include throughout the entire article. Some items simply require explicit statements, such as identifying the article as an RCT in the title (Item 1a) or explicitly stating any changes to the primary and secondary outcomes in a trial after the commencement of the trial, along with the reasons for the changes (Item 6b). Other items require more detail, such as the exploration of the study's limitations and sources of potential bias, imprecision, and, if relevant, the multiplicity of analyses in the paper's discussion (Item 20). Of note, Item 6a, which requires that primary and secondary outcome measures are "completely defined" and pre-specified, is the only item directly relating to reporting practices for measures.

The most recent version of JARS (Appelbaum et al., 2018) was published in 2018 and, similar to the 2010 CONSORT Statement (Schulz et al., 2010), includes information recommended for inclusion in quantitative research manuscripts, from the abstract through the discussion section. In addition to requiring researchers to clearly report designated primary and secondary measures, regardless of whether they were included in the report, it requires authors to report any steps they took to "enhance the quality of measurements" (p. 2), including reporting interrater reliability. No other guidelines are given for the reporting of the psychometric evidence for the measures included in the study. These reporting guidelines are important, as they help psychologists communicate in a standardized way about their procedures; however, the standards provide little guidance about reporting practices for measurement.

Though the explicit reporting of reliability and validity evidence for measures has not yet been included in CONSORT (Schulz et al., 2010) or JARS (Appelbaum et al., 2018), reliability and validity evidence are key components to rigorous study design and transparent reporting practices. The conclusions in psychological research are only as good as the measures used in the research, and psychologists may simply be using some measures because the measures are considered the standard in the field. For example, the Beck Depression Inventory, Second edition (BDI-II), is a popular measure for depression, with 10 of 14 studies identifying it as a tool used to a high degree in training and practice, a valuable clinical tool, and an instrument of choice in the assessment of mood disorders (Piotrowski, 2018). It is often assumed that these commonly used measures, such as the BDI-II, are only used so frequently due to the preponderance of psychometric evidence in their favor, though such evidence may not be commonly reported. Though such measures might have high reliability and validity, as is commonly assumed, without reporting the reliability or validity evidence of these measures, readers are expected to posit faith in the wisdom of the researchers' selection of measures. If, however, this assumption is false, the practice of using such "gold standard" measures with weak psychometric evidence would almost certainly result in problematic research that is not replicable or reflective of the real effects the researchers are attempting to study, essentially rendering such research useless.

To our knowledge, there has not been any systematic review of measurement reporting practices in psychotherapy clinical trials. As the first step in evaluating and improving measurement reporting is raising awareness of current measurement reporting practices, the primary purpose of this study is to review and evaluate the frequency with which researchers report and replicate reliability and validity evidence of the measures used. We will focus on psychotherapy research utilizing RCTs. Specifically, we will review what researchers report with respect to reliability and validity evidence of the measures used in RCTs published in the *Journal of Consulting and Clinical Psychology (JCCP)* in four different years. The JCCP is considered a leading journal in clinical psychology, with an impact factor of 4.54. It has regularly required studies published within it to comply with current reporting guidelines and currently requires RCTs to comply with the JARS (Appelbaum et al., 2018) reporting guidelines, so we expect the reliability and validity reporting practices of the research published in JCCP to be a valid representation of such reporting practices within the field of psychology. In addition, if reporting practices in early years are not exemplary, we expect that the implementation of reporting guidelines will improve the reporting and methodological rigor of studies published more recently as compared to earlier years; thus, our study will also focus on comparing the frequency with which reliability and validity evidence was reported in JCCP RCTs in 2018 as compared to 1994, 2002, and 2010. The first widely implemented reporting standard (CONSORT; Schulz et al., 2010) was published in 1996; our chosen years will sample RCTs prior to the first publication of CONSORT (Schulz et al., 2010), along with sampling RCTs six to nine years after its 1996 publication and each of its subsequent revisions in 2001 and 2010. A secondary aim of this study is to offer recommendations for improving the reporting practices within psychology so as to strengthen research findings and their replicability within psychology.

Method

Sampling and Procedures

We sampled randomized trials from four years of the Journal of Consulting and Clinical Psychology (JCCP): 1994, 2002, 2010, and 2018. After searching for all articles published in JCCP during these four years, we selected up to 15 articles from each year. For years with more than 15 RCT articles in a given year (2002, 2010, and 2018), we randomly sampled 15. Any JCCP articles in the given year without an RCT design were excluded from the study.

Our study methods and analyses are preregistered on the Open Science Framework (OSF), a free website designed to make the research process more transparent and reproducible. All data are stored on OSF and accessible to the public. The primary researchers have access to edit the data, and a full history of edits to the data and the analyses are publicly displayed on OSF. The link to our study is https://osf.io/rbz9t/?view_only=f9783f3340b64bc3afe021d6ef6f14ac.

Coding of Articles

All studies were coded by two graduate students. When there were discrepancies, the coders met and resolved discrepancies. Any discrepancies that could not be resolved between the two coders were resolved during a discussion with the thesis supervisor (Scott Baldwin).

Following the procedures outlined in the coding manual (included in the appendix), we created study-level codes and measure-level codes. Regarding study-level codes, we coded the year of publication, the chronological location of the RCT study in our sample from that year (e.g., 3 = the third study in that year), and the study's sample size, as recorded in the "Methods" section. Next, we coded whether the study reported one or more primary outcomes (1 = yes, 0 = no), the number of primary outcomes identified, whether the study identified one or more primary measures (1 = yes, 0 = no), and the number of primary measures identified. Primary outcomes are designated by the researcher as the most important outcome(s) (e.g., blood alcohol level, IQ score) among all outcomes examined, while primary measures are considered to be the most important measure(s) of one or more primary outcomes. In addition, we coded the total number of measures reported in the "Measures section" and the total number of measures reported in the study's tables. As the primary outcomes are typically stated in the last page of the "Introduction" section and the information on measures is typically explicated in the "Measures" section, we limited our search for this information to these sections of the sampled studies.

For each measure in the studies sampled, we coded for reliability and validity evidence as it is reported in the "Measures" section (or the equivalent section in the "Methods" section if no "Measures" section could be found). We focused on the "Measures" section because that is where the psychometric properties of the study's measures are typically reported. We created codes for the year, the chronological location of the study in our sample from that year as explained above, and the chronological location of the measure within the study (e.g., 2 = the second measure listed). We coded for whether subscales of a measure are treated as independent measures (1 = yes, 0 = no), and whenever this occurred, we treated each subscale as its own measure, coding for reliability and validity evidence for each subscale. In addition, we coded for whether the measure consisted of a selection of items from a larger measure or subscale (1 = yes, 0 = no). We also coded whether the measure was identified as a primary measure (1 = yes, 0 = no) and whether the measure was created or altered by the authors primarily for use in the given study (0 = not created or altered, 1 = created by author(s), 2 = altered by author(s), 3 =

unknown/unclear). In addition, we recorded the construct being measured and the method (e.g., self-report, formal assessment, physical/biological measure) employed by the measure.

Because we expected that the most common reliability statistic reported would be Cronbach's alpha, we coded for the presence of alpha calculated from the given study's data (1 = yes, 0 = no), the presence of another reliability statistic calculated in the given study (1 = yes, 0 = no), the presence of citations reporting one or more reliability statistics from previous data (1 = yes, 0 = no), the presence of qualitative reporting of reliability evidence (1 = yes, 0 = no), and the number of studies cited to support reliability evidence. We repeated this coding procedure for the internal consistency, test-retest, and interrater reliability evidence for every reported measure in the sampled studies.

Similarly, we coded for the presence of a validity statistic calculated from the given study's data (1 = yes, 0 = no), the presence of citations reporting one or more validity statistics from previous data (1 = yes, 0 = no), the presence of qualitative reporting of validity evidence (1 = yes, 0 = no), and the number of studies cited to support validity evidence. We repeated this coding procedure for the construct, factorial, convergent, discriminant, and predictive validity evidence for every reported measure in the sampled studies.

During the coding process, we encountered a study that only employed a subscale of a measure, yet reported psychometric evidence for the entire measure and not for the subscale. In anticipation of encountering more studies that employed this practice, we added a third option to all reliability and validity evidence codes (2 = evidence for the entire measure, but not specifically for the subscale/portion researchers are using).

Interrater reliability

To assess the interrater reliability of the coding procedure described above, two graduate students independently coded ten JCCP RCT studies from 2017. For categorical variables, percentage agreement ranged from 80% to 100%, with a median of 100%; kappa ranged from 0.62 to 1.00, with a median of 1.00; and for continuous variables, the correlation between raters ranged from $r = 0.75$ to 1.00, with a median of 1.00. All variables were sufficiently reliable to proceed with the analysis.

Data Analysis

We used R (R Core Team, 2016) to randomly sample 15 articles from each year in which more than 15 RCTs meeting our inclusion criteria were present; these were 2002, 2010, and 2018. 1994 had exactly 15 RCTs meeting our inclusion criteria. After randomly selecting articles from all four years, we encountered 12 articles in 2010 and one article in 2002 that either did not meet sampling criteria (and thus were excluded from analysis) or stated in the "Methods" section that more information about participants and/or measures were recorded in an earlier study. Consequently, these articles were excluded from analysis, as they would not be expected to report the same detail of information on measures and participants as articles reporting on this data for the first time. The random sampling process was repeated for all articles not yet coded from that year until 15 articles meeting all inclusion criteria had been selected. All statistical analyses were run in Stata 15 (StataCorp, 2017). The following PRISMA flow diagrams demonstrate our sampling procedures.

PRISMA 2009 Flow Diagram: 1994 Sample

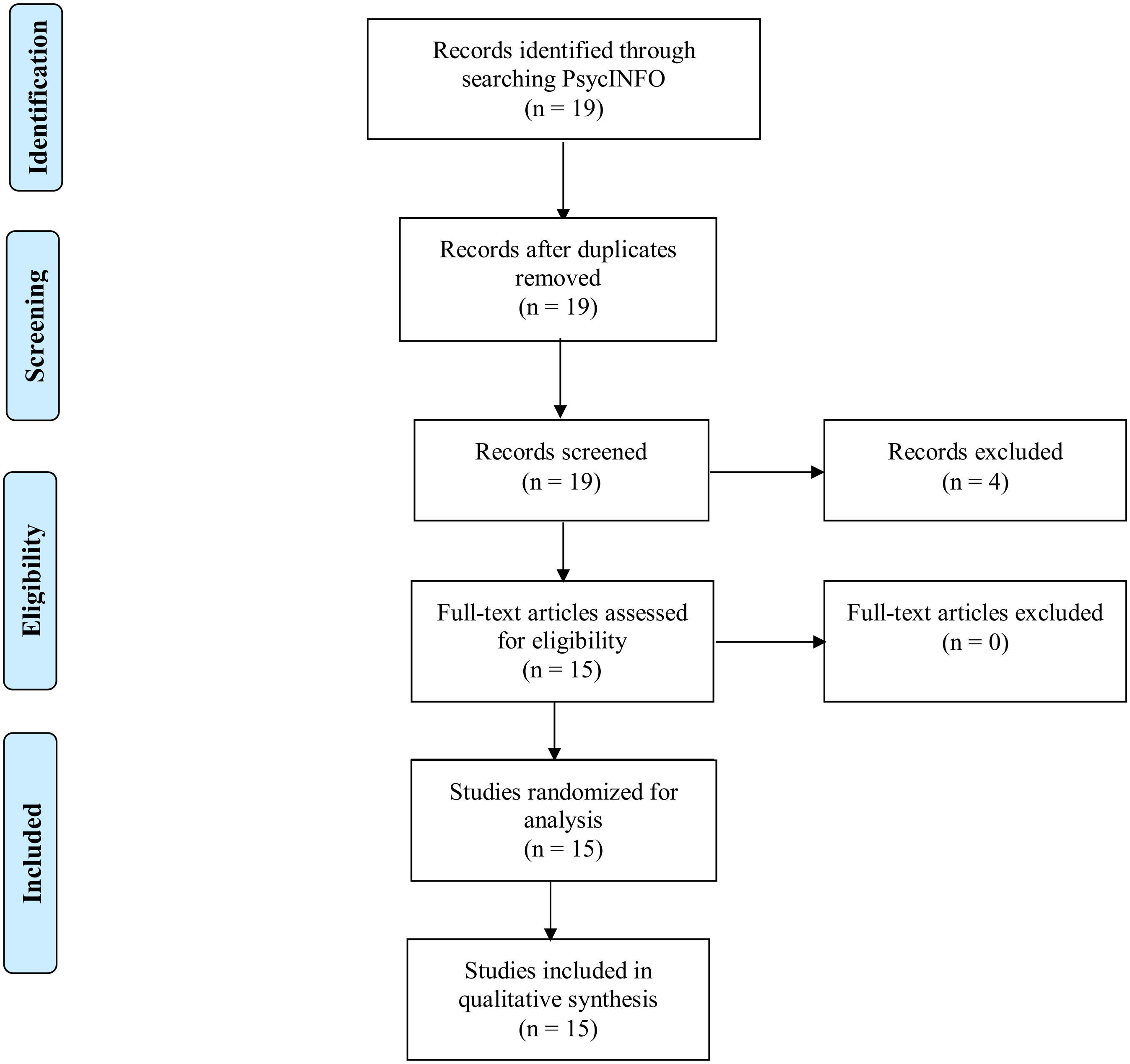
Identification
Screening
Eligibility
Included
Records identified through searching PsycINFO (n = 19)
Records after duplicates removed (n = 19)
Records screened (n = 19)
Records excluded (n = 4)
Full-text articles assessed for eligibility (n = 15)
Full-text articles excluded (n = 0)
Studies randomized for analysis (n = 15)
Studies included in qualitative synthesis (n = 15)

Figure 1. PRISMA 2009 flow diagram: 1994 sample.

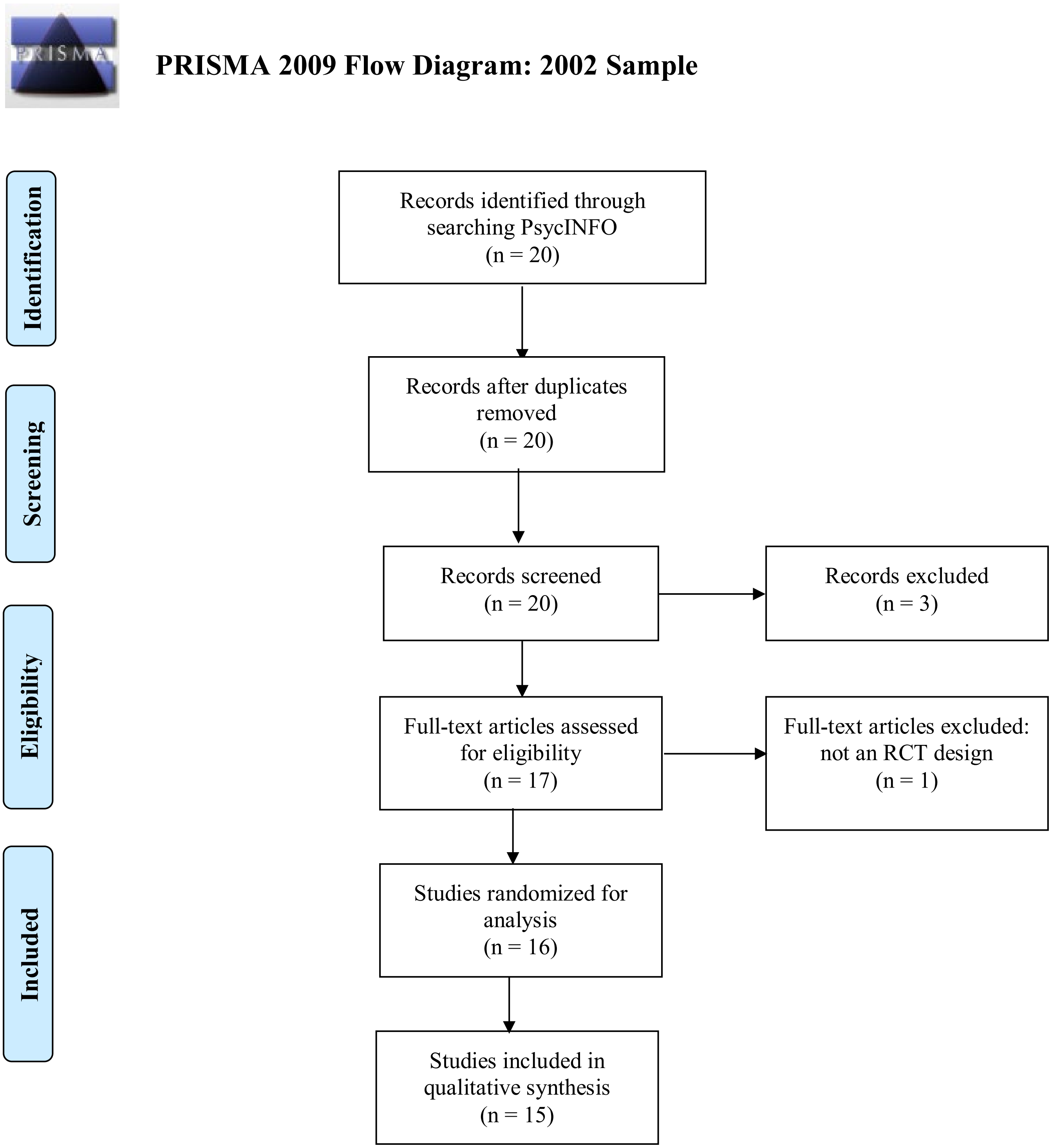
PRISMA
PRISMA 2009 Flow Diagram: 2002 Sample
Identification
Screening
Eligibility
Included
Records identified through searching PsycINFO (n = 20)
Records after duplicates removed (n = 20)
Records screened (n = 20)
Records excluded (n = 3)
Full-text articles assessed for eligibility (n = 17)
Full-text articles excluded: not an RCT design (n = 1)
Studies randomized for analysis (n = 16)
Studies included in qualitative synthesis (n = 15)

Figure 2. PRISMA 2009 flow diagram: 2002 sample.

PRISMA 2009 Flow Diagram: 2010 Sample

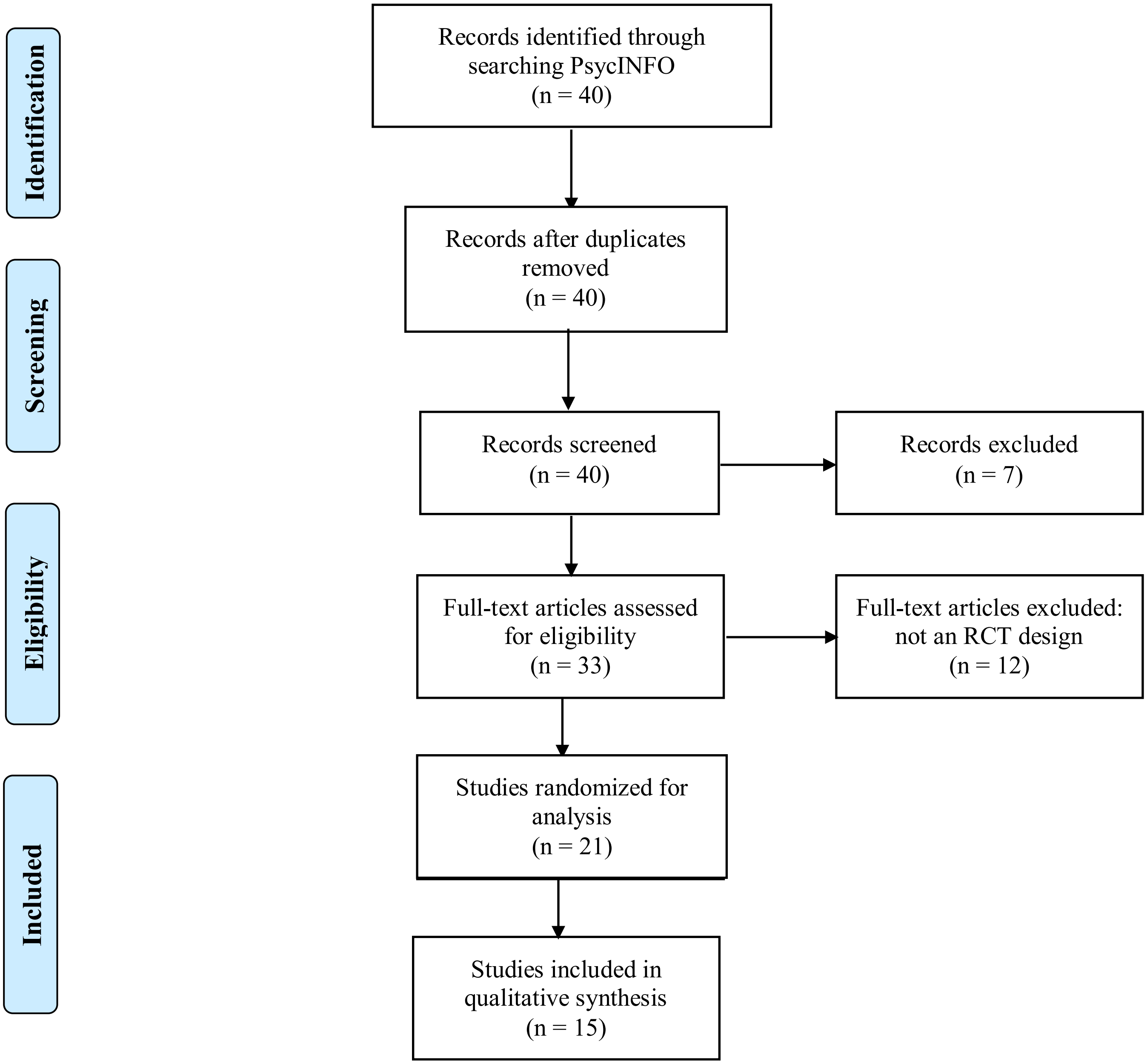
Identification
Screening
Eligibility
Included
Records identified through searching PsycINFO (n = 40)
Records after duplicates removed (n = 40)
Records screened (n = 40)
Records excluded (n = 7)
Full-text articles assessed for eligibility (n = 33)
Full-text articles excluded: not an RCT design (n = 12)
Studies randomized for analysis (n = 21)
Studies included in qualitative synthesis (n = 15)

Figure 3. PRISMA 2009 flow diagram: 2010 sample.

PRISMA 2009 Flow Diagram: 2018 Sample

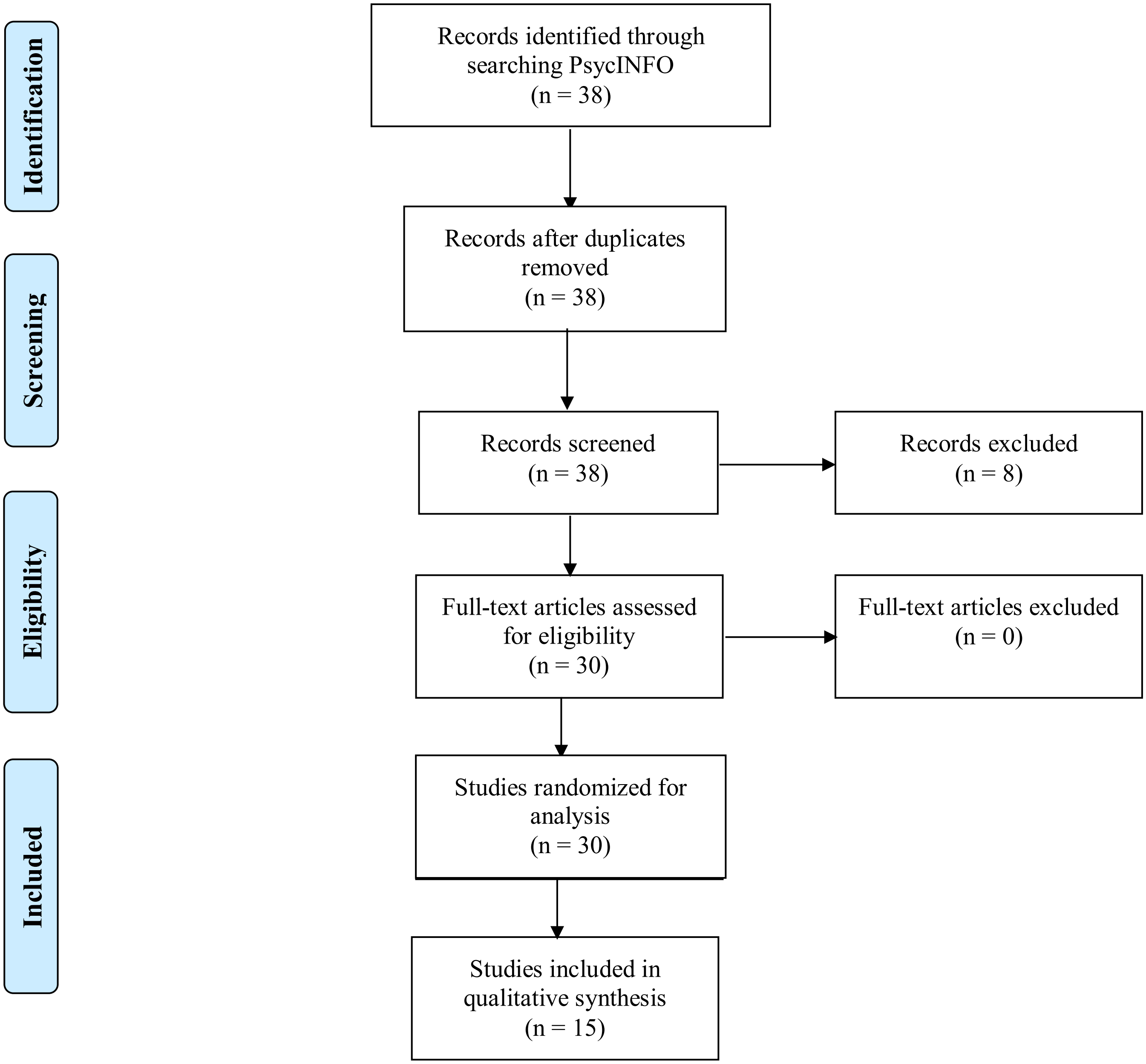
Identification
Records identified through searching PsycINFO (n = 38)
Screening
Records after duplicates removed (n = 38)
Records screened (n = 38)
Records excluded (n = 8)
Eligibility
Full-text articles assessed for eligibility (n = 30)
Full-text articles excluded (n = 0)
Included
Studies randomized for analysis (n = 30)
Studies included in qualitative synthesis (n = 15)

Figure 4. PRISMA 2009 flow diagram: 2018 sample.

Results

Description of Studies and Measures

On average, we coded 7.5 (standard deviation [SD] = 4.4) measures per article, with a total of 450 measures across all studies. Self-report measures were the most common type of measure used, with 37.6% (169) of measures designated by the authors as self-report measures; 31.3% (141) of the measures did not report enough information to deduce the measure type. Of the 450 measures, 2.0% (9) were created by the study authors for use in the study, and 4.9% (22) were altered by the study authors. 12.4% (56) of the measures appeared to have been created or altered, but it was unclear whether they had been altered by study authors for use in the coded study or if they had been used in their altered forms in other studies. Table 1 displays the percent of studies identifying primary outcomes and primary measures for each year of articles sampled. Primary outcomes are the researcher-designated most important outcome(s) (e.g., blood alcohol level, IQ score) among all outcomes examined, while primary measures are considered to be the most important measure(s) of one or more primary outcomes.

Table 1
Percent Primary Outcomes and Measures Identified by Year Published

Year	Number of Studies	Mean Number of Measures per Study	Primary Outcomes Identified	Primary Measures Identified
1994	15	9.6	0.0	6.7
2002	15	7.7	6.7	6.7
2010	15	5.7	26.7	33.3
2018	15	7.0	66.7	60.0

Reliability and Validity Evidence Reported

Table 2 shows the percent of measures where a study reported any kind of reliability evidence and the percent of measures where a study reported any kind of validity evidence for each year of articles sampled. This includes reliability and validity evidence cited from previous studies (including those not reporting a statistic), along with evidence calculated from the given study's data (current reliability/validity evidence). Internal consistency reliability was the most frequent type of reliability evidence reported, with internal consistency reliability evidence reported on 25.6% (115) of all measures and 52.4% (55) of 2018 measures. Convergent validity was the most frequent type of validity evidence reported, with convergent validity evidence reported on 4.0% (18) of all measures and 6.7% (7) of 2018 measures.

Table 2
Percent Measures Reporting Reliability and Validity Evidence by Year Published

Year	Number of Measures	Reliability Evidence	Validity Evidence
1994	144	18.1	3.5
2002	116	23.3	4.3
2010	85	47.1	14.1
2018	105	58.1	12.4

Table 3 shows the percent of measures reporting current reliability evidence, and the percent of measures reporting current validity evidence was calculated from the given study's data each year. Internal consistency reliability was the most common type of current reliability evidence reported, with 20.7% (93) of all measures and 48.6% (51) of 2018 measures reporting internal consistency reliability calculated from the given study. Therefore, almost half of all measures reporting current internal consistency reliability evidence were from 2018 studies. Convergent validity was the most common type of current validity evidence, with 0.9% (4) of all measures and 2.9% (3) of 2018 measures reporting convergent validity calculated from the given study. Therefore, all but one of the measures reporting convergent validity evidence were from 2018 studies. Of the 154 measures for which reliability evidence was reported, 15.6% (24) used a descriptive term with no accompanying statistic to report some form of reliability evidence. Of the 35 measures for which validity evidence was reported, 71.4% (25) used a descriptive term with no accompanying statistic to report some form of validity evidence.

Table 3

Percent Measures Calculating Reliability and Validity Evidence by Year Published

Year	Number of Measures	Reliability Evidence	Validity Evidence
1994	144	11.8	0.0
2002	116	19.8	0.0
2010	85	30.6	1.2
2018	105	51.4	2.9

Regression Analysis

Reliability Evidence

Table 4 displays the unstandardized regression coefficients for the percent of measures reporting reliability evidence by year published. Articles published in 2018 reported significantly more reliability evidence for their measures than articles published in 1994 (*b* = - 0.40 p < 0.001; 95% Confidence Interval [CI] = [- 0.51, - 0.29]) or 2002 (*b* = - 0.35, p < 0.001; 95% CI = [- 0.47, - 0.23]). For example, an article published in 1994 reported reliability evidence for 40% fewer of its measures, on average, than the average article published in 2018. Articles published in 2010 did not differ significantly from those published in 2018 with respect to the percent of measures for which reliability evidence was reported (*b* = - 0.11, p < 0.09; 95% CI = [- 0.24, 0.02]).

Table 4

Regression Analysis Predicting Percent Reliability Evidence by Year Published

Variable	B	SE b	p	95% Confidence Interval	
				Lower	Upper
Constant	0.58	0.04	< 0.001	0.50	0.67
1994	- 0.40	0.06	< 0.001	- 0.51	- 0.29
2002	- 0.35	0.06	< 0.001	- 0.47	- 0.23
2010	- 0.11	0.07	0.09	- 0.24	0.02

Notes. R^2=0.12

Validity Evidence

Table 5 displays the unstandardized regression coefficients for the percent of measures reporting validity evidence by year published. Articles published in 2018 reported significantly more validity evidence for their measures than articles published in 1994 (*b* = - 0.09, p = 0.01; 95% CI = [- 0.16, - 0.02]) or 2002 (*b* = - 0.08, p = 0.02; 95% CI = [- 0.15, - 0.01]). For example, an article published in 1994 reported validity evidence for 9% fewer of its measures, on average, than the average article published in 2018. Articles published in 2010 did not differ significantly from those published in 2018 with respect to the percent of measures for which validity evidence was reported (*b* = 0.02, p < 0.65; 95% CI = [- 0.06, 0.09]).

Table 5

Regression Analysis Predicting Percent Validity Evidence by Year Published

Variable	B	SE b	p	95% Confidence Interval	
				Lower	Upper
Constant	0.12	0.03	< 0.001	0.07	0.17
1994	- 0.09	0.03	0.01	- 0.16	- 0.02
2002	- 0.08	0.04	0.02	- 0.15	- 0.01
2010	0.02	0.04	0.65	- 0.06	0.09

Notes. R^2=0.03

Discussion

The primary purpose of this study is to review and evaluate the frequency with which researchers report and replicate reliability and validity evidence of the measures used. A secondary aim of this study is to offer recommendations for improving the reporting practices within psychology so as to strengthen research findings and their replicability within psychology.

Reporting of Reliability Evidence

The vast majority of psychological research measures latent constructs. One challenge with measuring a latent variable is measuring it with sufficient reliability. If a measure of a construct is not sufficiently reliable, it is difficult, if not impossible, to detect true differences between groups, as the error variability will overshadow the true variability. The best measures are consistent in multiple ways, including demonstrating consistency within a given sample (internal consistency reliability; ICR), across time (test-retest reliability), and across raters (if the measure is coded; interrater reliability). ICR evidence was the most commonly reported reliability evidence, yet it was only reported for 25.6% of the measures sampled. In 2018, ICR evidence was reported for 52.4% of measures, suggesting that though psychological articles are far from reporting ICR evidence for every measure, ICR evidence reporting practices have improved over time, possibly due to the implementation of reporting standards such as CONSORT and JARS.

In addition, the reliability statistic generated for a measure in one study may differ due to changes in sample characteristics such as age, gender, or race. Because of this, calculating reliability statistics of a measure in the current sample is a crucial step to determining whether the measure produces reliable results in a given study. ICR evidence was the most common type of current reliability evidence reported, and yet it was only reported for 20.7% of the measures in our sample. In 2018, current ICR evidence was reported for 48.6% of measures. Therefore, though psychological articles are far from calculating ICR for every measure in every study, current ICR evidence reporting practices have improved over time.

Although studies published in 2018 reported significantly more reliability evidence for their measures than those published in 1994 or 2002, still only slightly more than half of the measures from studies published in 2018 reported some form of reliability evidence. Thus, almost half of all measures used in 2018 JCCP studies do not report reliability evidence of any kind, despite JCCP requiring RCTs to follow the JARS (Appelbaum et al., 2018) guidelines. These studies may have reported a general statement regarding the psychometric properties of the measure; this practice will be discussed in further depth below. Without reporting reliability evidence, it is difficult to know whether study results reflect true patterns in the data or patterns generated by error variance.

Reporting of Validity Evidence

Another challenge in psychological research is creating valid measures. As with reliability, the best measures demonstrate multiple forms of validity, including the extent to which a measure captures the construct it is designed to measure (construct validity), to which a measure's items reflect latent factors (factorial validity), to which a measure generates results consistent with another way to measure the construct (convergent validity), to which a measure generates results that differ from measures of unrelated constructs (discriminant validity), and the extent to which a measure is able to predict results related to the construct (predictive validity). Although studies published in 2018 reported significantly more validity evidence for their measures than those published in 1994 or 2002, still only about one of every eight measures from studies published in 2018 reported some form of validity evidence, and less than 3% of

2018 measures calculated a validity statistic from the current sample. Without reporting and replicating validity evidence for our measures, researchers cannot draw conclusions about whether the scores reflected in the study are a valid representation of the construct of interest. If a measure is not valid, the conclusions drawn from the measure's data are likely to be invalid as well.

Researchers may not report validity evidence for a variety of reasons. For example, they may rely on the measures traditionally used in their field of study, assuming the widespread use of such measures is founded on evidence of its validity or a basis for validity evidence. Many measures are face valid, and because they appear to be asking about the construct of interest, researchers may consider such face validity as necessary and sufficient evidence that the measure is accurately assessing what it was designed to measure. Similarly, researchers may hold the misconception that reliability is necessary and sufficient evidence for validity (e.g., that people who consistently report a high score on a measure of anxiety must be high in anxiety). In addition, validity is more difficult to measure, as no widely accepted "validity statistic" exists; therefore, it may be more difficult to measure and to locate instances of validity evidence in past research. Perhaps the most important reason researchers may not report validity evidence is because few editors and reviewers mandate researchers to demonstrate that the study's measures are valid for their use. Regardless of the underlying reason, researchers fail to report validity evidence, the current dearth of studies reporting validity evidence for their measures indicates the need for improvement.

Impact of Implementing Reporting Standards

Over time, psychological research practices regarding reporting reliability and validity evidence has improved, though it did not change significantly between 2010 and 2018. Along with the replication crisis and subsequent focus on improving the rigor of psychological research, the implementation of reporting standards such as CONSORT (Schulz et al., 2010) is likely responsible for some of the improvement in psychometric reporting practices. It appears, however, that the 2010 update of CONSORT (Schulz et al., 2010), along with the implementation of the first publication of JARS (Appelbaum et al., 2018) in 2008, may not have improved psychometric reporting of measures within psychological research, as 2018 did not significantly differ from 2010 in the percent of measures reporting reliability and validity evidence. Such reporting standards are vital to incentivizing better reporting practices. However, a substantial percentage of measures do not have reliability and validity evidence reported in studies published in a top psychology journal in 2018. We recommend some simple changes that researchers can implement to increase the transparency surrounding psychometric reporting of the measures they use.

Recommendations

First, researchers should be required to report more than ICR evidence for the measures included in the study. Although ICR helps us understand whether a given sample demonstrated consistent patterns of responding, it does not tell us anything about the measure's reliability across any other dimension (e.g., across time). In addition, reliability is necessary but not sufficient for validity; i.e., a measure must be reliable to be valid, but reliability says nothing about whether the measure is capturing the construct it was designed to measure. In addition to reporting ICR, for any measures that have been used across multiple timepoints or that require coding, we recommend reporting test-retest or interrater reliability, respectively, for those measures.

In addition, we recommend reporting validity evidence for all measures included in the study. Ideally, researchers would report multiple forms of validity that have been tested in the past; however, if every study reported just one form of validity for their measures, this would greatly improve psychometric reporting practices. Most commonly used measures have published validation studies; researchers are doing a disservice to the field of psychological research if they fail to report the validity evidence that has already been generated for their measures.

Calculate Current Psychometrics

A commonly overlooked limitation to reliability and validity evidence generated in previous studies is its generalizability to samples, unlike the original sample. As reliability and validity statistics are calculated from a finite sample, the only way for researchers to know whether the reliability and validity of the measure apply to their sample is to attempt to replicate all possible reliability and validity statistics with their own data. Although over half of the measures published in our sampled studies from 2018 reported ICR calculated from the given study's data, this still means that almost half of the measures in our sample did not report an ICR statistic generated from the current data. Calculating ICR does not require any additional data collection, nor does it require complex statistical analyses; therefore, all researchers should be expected to report an ICR statistic calculated from the current data. Similarly, studies employing measures across multiple time points should be expected to calculate test-retest reliability for their current data, and studies with measures that require an independent coder should have a second coder for at least a portion of the data in order to calculate interrater reliability for the study's data. Such calculations are a type of integrity check: a way for researchers to determine whether their measures are behaving consistently throughout their study.

We propose that researchers also calculate all possible validity statistics from the data they collect. Most studies we coded included at least two measures of a construct, which gives researchers enough information to calculate convergent validity for their study. Similarly, we encourage researchers to report any other forms of validity evidence for which they have collected data. The types of validity evidence researchers are most likely to be able to calculate from their studies with minimal changes to study design include construct and discriminant validity evidence.

Clarify Writing

We propose that researchers should clarify their writing, making primary outcomes and primary measures more explicit. Only two-thirds of articles published in 2018 clearly stated primary outcomes and/or measures, despite the 2010 Consort Statement checklist (Schulz et al., 2010) mandating much more rigorous reporting of primary outcome measures (p. 699) than the current study's coding manual of primary outcome measures. Simply adding a sentence that clearly designates which outcomes researchers consider to be of primary interest in answering their main research question eliminates confusion about the main focus of the study. Identifying primary outcomes combined with pre-registration of hypotheses and analyses will help improve the transparency of clinical trials.

Similarly, by explicitly designating which measures are considered primary measures, researchers will eliminate confusion about which measures are directly measuring primary outcomes. Such a clear designation of primary and secondary outcomes and measures may also aid researchers in using fewer measures. On average, our sampled studies reported 7.5 measures per article. Type I error rate often increases as more measures are utilized. Therefore, we propose that researchers use no more than two to three measures to address each of the primary outcomes and that researchers minimize the number of secondary outcomes they include in their studies. In addition, we propose that all measures included in the study clearly relate to one primary or secondary outcome, as required by step 6a of the 2010 CONSORT Statement checklist (Schulz et al., 2010, p. 699).

We also encourage researchers to explicitly report secondary outcomes and measures to increase the clarity of their writing and research design. Similar simple changes, such as explicitly stating the type of measure (e.g., self-report, parent report, structured interview; almost one-third of articles did not clearly state this) and whether measures were created or altered by the author for use in the current study, would greatly increase the transparency of researcher practices and clarify their writing.

Report Specifics of Psychometric Evidence

Last, when reporting reliability and validity statistics, we encourage researchers to include the statistics whenever possible. Many studies gave qualitative statements about the reliability and validity evidence for a study (e.g., "[Measure X] has demonstrated acceptable internal consistency reliability"); while such a statement is useful in helping others interpret reliability and validity statistics, it should not replace the reporting of the actual statistic. In cases where multiple studies generate a range of reliability or validity statistics, we encourage researchers to report the entire range found across studies. By including the specific statistic in reporting psychometric evidence for measures, consumers of research will be able to get a more precise estimate of the overall reliability and validity of a given measure. Further, reporting of statistics assists with peer review, both pre- and post-publication.

Similarly, we encourage researchers to report specific types of reliability and validity evidence for each of their measures. 9.8% (44) of measures reported "good psychometric properties"; such a vague statement does not give the reader any idea of the types of reliability and validity evidence that have been generated for the measure. It gives the impression of reliability and validity for the measure in such vague language as to leave everything to the reader's imagination. Instead, we encourage researchers to give the specific types of reliability and validity evidence for the measure, along with the statistics, as mentioned above.

Conclusion

In conclusion, although the implementation of reporting practices such as CONSORT (Schulz et al., 2010) have increased the frequency of reporting reliability and validity evidence, the current reporting frequency (18.1 – 58.1% for reliability evidence and 3.5 – 12.4% for validity evidence) is not sufficient. Without a clear picture of the reliability and validity of our measures, we cannot get a clear picture of the accuracy of our conclusions in scientific research. We recommend simple changes that impose a minimal burden on researchers; yet these changes, if implemented, have great potential to move us forward in the pursuit of truth.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, *63*(1), 32–50. https://doi.org/10.1037/0003-066X.63.1.32

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

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

Clayson, P. E., & Miller, G. A. (2017). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. *International Journal of Psychophysiology*, *111*, 57–67. https://doi.org/10.1016/j.ijpsycho.2016.09.005

Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. Submitted to jour *Advances in Methods and Practices in Psychological Science* Nov 2019. https://doi.org/10.31234/osf.io/hs7wm

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders*, *172*, 96–102. https://doi.org/10.1016/j.jad.2014.10.010

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, *12*(4), 660–664. https://doi.org/10.1177/1745691616687745

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

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

Murray, S. B., Compte, E. J., Quintana, D. S., Mitchison, D., Griffiths, S., & Nagata, J. M. (2019). Registration, reporting, and replication in clinical trials: The case of anorexia nervosa. *International Journal of Eating Disorders*, eat.23187. https://doi.org/10.1002/eat.23187

Naragon-Gainey, K., McMahon, T. P., & Park, J. (2018). The contributions of affective traits and emotion regulation to internalizing disorders: Current state of the literature and measurement challenges. *American Psychologist*, *73*(9), 1175–1186. https://doi.org/10.1037/amp0000371

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, *31*(3). https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science

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

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Piotrowski, C. (2018). The status of the Beck inventories (BDI, BAI) in psychology training and practice: A major shift in clinical acceptance. *Journal of Applied Biobehavioral Research*, *23*(3), e12112. https://doi.org/10.1111/jabr.12112

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https:www.R-project.org/

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, *340*(mar23 1), c332–c332. https://doi.org/10.1136/bmj.c332

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

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A.

(2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

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

Smith, G.T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 300–308. https://doi-org.erl.lib.byu.edu/10.1037/1040-3590.7.3.300

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886–899. https://doi.org/10.1177/1745691615609918

StataCorp. (2017). *Stata Statistical Software: Release 15.* StataCorp LLC.

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*(5), 742–756. https://doi.org/10.1177/1745691617690042

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, *3*(1), 1. https://doi.org/10.1525/collabra.74

Wood, J. M., Garb, H. N., Lilienfeld, S. O., & Nezworski, M. T. (2001). Clinical assessment. *Annual Review of Psychology*, *53*(1), 519-543. https://doi-org.erl.lib.byu.edu/10.1146/annurev.psych.53.100901.135136