aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2014, Vol. 9(6) 626–640
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691614551749
pps.sagepub.com
SAGE

A Duty to Describe: Better the Devil You Know Than the Devil You Don't

Sacha D. Brown[1], David Furrow[2], Daniel F. Hill[1], Jonathon C. Gable[1], Liam P. Porter[3], and W. Jake Jacobs[1]
[1]University of Arizona, [2]Mount Saint Vincent University, and [3]Boston University

Abstract

Although many researchers have discussed replication as a means to facilitate self-correcting science, in this article, we identify meta-analyses and evaluating the validity of correlational and causal inferences as additional processes crucial to self-correction. We argue that researchers have a duty to describe sampling decisions they make; without such descriptions, self-correction becomes difficult, if not impossible. We developed the Replicability and Meta-Analytic Suitability Inventory (RAMSI) to evaluate the descriptive adequacy of a sample of studies taken from current psychological literature. Authors described only about 30% of the sampling decisions necessary for self-correcting science. We suggest that a modified RAMSI can be used by authors to guide their written reports and by reviewers to inform editorial recommendations. Finally, we claim that when researchers do not describe their sampling decisions, both readers and reviewers may assume that those decisions do not matter to the outcome of the study, do not affect inferences made from the research findings, do not inhibit inclusion in meta-analyses, and do not inhibit replicability of the study. If these assumptions are in error, as they often are, and the neglected decisions are relevant, then the neglect may create a good deal of mischief in the field.

Keywords

self-correction, direct replication, systematic replication, conceptual replication, meta-analysis, publication, methodology

> We should value replication more than we do, treasure it even. We were all routinely taught the value of replication in our first research methods course, but it seems some have forgotten the lesson. (Roediger, 2012, para. 6)

Many authors have made general recommendations for increased sharing of study details as a means of addressing problems of direct[1] (Schmidt, 2009; Sidman, 1960), systematic[2] (Sidman, 1960), and conceptual[3] (Schmidt, 2009) replication (e.g., Asendorpf et al., 2013; Eich, 2014; Grant et al., 2013; LeBel et al., 2013; LeBel & Peters, 2011). We claim these calls do not go far enough. Our proposal builds on previous suggestions in several ways. First, we outline the information necessary for adequate correlational and causal inferences, replication, and inclusion in meta-analyses from a view grounded in Mill's (1843) Canons—constrained in terms of what (a) a researcher can realistically attend to or be in control of and (b) has been shown to influence the outcomes of several studies. Second, we place the responsibility for reporting of sampling decisions on authors and journals equally and assert these descriptions should be available with published studies. Third, we address how researcher failures to describe sampling decisions may damage several aspects of scientific self-correction beyond replication, including meta-analysis and inferential validity of a study. Finally, we document that a range of psychological journals underdescribe many sampling decisions, not just purely methodological ones, and we argue that they matter.

That science self-corrects is a truism both old and new (e.g., Bordens & Abbott, 2008; Cohen & Nagel, 1934;

Corresponding Author:
Sacha D. Brown, Department of Psychology, University of Arizona, 1503 East University Blvd., Tucson, AZ 85721
E-mail: sdbrown@email.arizona.edu

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

Goodwin & Goodwin, 2013). Scientists count this self-correcting "way of knowing" as unique because its strong empirical ties allow reproducibility and, through that, self-correction. Among the old, Cohen and Nagel (1934) claimed, "Other methods . . . are all inflexible, that is, none . . . can admit that it will lead us into error. Hence, none of them can make provision for correcting its own results" (p. 195). Among the new, Lilienfeld, Lynn, Namy, and Woolf (2009) named "lack of self-correction" as the second deadly sin of pseudoscience (pp. 45–46). Replication is a major mechanism for self-correction, and without a willingness to share complete description of our studies with others, "the scientific enterprise grinds to a screeching halt, because research progress hinges on the ability to evaluate other investigators' findings objectively" (Lilienfeld et al., 2009, p. 29).

Recent special sections and issues of journals reflect a growing apprehension over the role of direct, systematic, and conceptual replications, as well as meta-analyses, suggesting that science may be self-correcting more in theory than in practice.[4] Ioannidis (2005) has been at the forefront of researchers identifying factors interfering with self-correction. He has claimed that journal editors selectively publish positive findings and discriminate against study replications, permitting errors in data and theory to enjoy a long half-life (see also Ferguson & Brannick, 2012; Ioannidis, 2008, 2012; Shadish, Doherty, & Montgomery, 1989; Stroebe & Strack, 2014). We contend there are other equally important, yet relatively unexplored, problems.

One neglected problem is a failure of original authors to report sampling decisions adequately. This creates a three-fold problem for self-correcting science: (a) it forces replicating researchers to guess what sampling decisions an original researcher made; (b) it leaves meta-analytic researchers unable to include all eligible studies, especially when testing potential moderators; and (c) it leaves replicating researchers, meta-analysts, editors, and readers unable to accurately assess the validity of inferences made by the researcher of a study. To understand why this is, we need to understand what exactly sampling decisions are and how they can change inferences made.

What Are Sampling Decisions?

Although researchers in the field often think that a sampling decision, taking a small sample from a larger population, refers to the number and type of participants chosen, the majority of decisions that researchers make involve sampling decisions. A researcher must choose samples (often nonrandom samples) from several distinct populations: most obviously, the populations of participants, independent (predictor) variables, and dependent (outcome) variables, as well as measures of them. Asendorpf et al. (2013) pointed out that sampling decisions extend to experimental situations and time points relevant to a study's design. Even more broadly, researchers must also sample from populations of experimenters, assessors, physical settings, measures, and available statistical analyses. We use the phrase *sampling decisions*[5] to refer to all of these sampling choices.

The same psychological judgment biases that plague humans in general may taint any sampling decision (Fiedler, 2011). As important, although many studies use large *N* samples for the participants, the remaining samples are small *N* (often $N = 1$). It is common to see studies run by one (undescribed) experimenter, under one (partially described) set of experimental settings, using a few operationalizations of an independent/predictor or dependent/outcome variable. The small *N*s involved in these samples make it increasingly likely that the study will produce an extreme outcome—an error (Wainer, 2007). Coupled with journals' tendencies to publish "hot" new results, biases in the literature become even more probable.

The Role of Sampling Decisions and Inferential Validity

In his classic, *System of Logic*, John Stuart Mill (1843) gave us a set of Canons allowing us to infer that a specific event caused or is related to another event with some confidence. Mill proposed these inductive principles as a way to regulate (and regularize) scientific inquiry.

We have included Table 1, which illustrates three of these principles, to remind us that the logic of our correlational and experimental designs rests on Mill's (1843) Canons. A failure to describe the sampling decisions used to instantiate the logic of a Canon limits or eliminates the ability to assess the inferential validity of a researcher's interpretation of a study's outcome, to conduct replications of any kind, and to include eligible studies in meta-analyses. Consequently, this jeopardizes both self-correction and the entire inferential enterprise of the science.

To illustrate this problem, consider the implications of a recent article appearing in *Nature Methods*. The authors reported profound and differential effects of the presence of male and female humans on the performance of the most common of subjects—rats and mice.

> [Exposure to] Male- [but not female-] related stimuli induced a robust physiological stress response that results in stress-induced analgesia. This effect could be replicated with T-shirts worn by men, bedding material from gonadally intact and unfamiliar male mammals, and presentation of compounds secreted from the human axilla. Experimenter sex can thus affect apparent baseline responses in behavioral testing. (Sorge et al., 2014, p. 629)

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

Table 1. Mill's (1843) Canons

Canon	Relevant research designs
Method of Agreement: If there are several examples of an observed phenomenon, And if these phenomena have one and only one preceding circumstance in common, Then that common event is the cause of the phenomenon.	Descriptive, correlational, quasi-experimental
Method of Difference: If two conditions are identical save for one circumstance, And if a phenomenon under investigation appears in the presence of the circumstance, And if a phenomenon under investigation does not appear in the absence of the circumstance, Then that circumstance is the effect, or cause, or an indispensable part of the cause, of the phenomenon.	Experimental (e.g., randomized groups design, matched groups design, small *N* designs)
Method of Concomitant Variations: If a phenomenon varies in any manner, And another phenomenon varies in a similar manner, Then these phenomena are connected through some fact of causation.	All of the above

Note: This table contains a summary of the logical syllogisms that constitute Mill's (1843) Canons that are relevant to our discussion: the Method of Agreement, the Method of Difference, and the Method of Concomitant Variations. Researchers in the field apply these methods to make inferences regarding the causal relations among phenomena (or events).

Researchers in behavioral or physiological psychology, physiology, and the neurosciences seldom describe the sex of the experimenters running their subjects or participants. Without that knowledge, we cannot know whether a basic premise of Mill's (1843) *Method of Difference* (that all conditions in an experiment are, in the beginning, identical) holds, making all causal inferences based on the data suspect. This fact has led some to claim, "Decades of science are going to be—perhaps not voided but certainly called into question" (Petri, 2014, para. 7).

The controversy surrounding the well-known priming study (Bargh, Chen, & Burrows, 1996) in which young participants walked more slowly after completing a scrambled-sentence task priming an elderly stereotype than participants in a control condition provides an example from literature in which human participants were used. Despite its fame, many failed to replicate the study's findings. In the original report, Bargh et al. (1996) did not describe all their procedures, including the fact that the experimenter who handed the priming or control task to participants was the same individual who packaged the materials (see Yong, 2012, for an extended discussion). In 2012, Doyen, Klein, Pinchon, and Cleeremans demonstrated the potential for this methodological "detail" to play a crucial role in the outcome of the original study. Bargh et al. inferred that participant priming caused the outcome, but others could not replicate the finding purportedly supporting this causal inference without this overlooked information. Researchers have spent more than a decade wondering what went awry in their replication attempts and even now continue to propose alternative inferences (e.g., Cesario, Plaks, & Higgins, 2006; Doyen et al., 2012; Hull, Slone, Metayer, & Matthews, 2002; Klatzky & Creswell, 2014; Pashler, Harris, & Coburn, 2011).[6]

As outlined earlier, to implement Mill's (1843) Canons, researchers make a large number of sampling decisions. They make judgments about which of these decisions are relevant, they control for factors that they believe are germane, and they ignore those that they intuit are not. Each choice is a sampling decision. Note that almost none of the resulting samples are taken randomly, instead they are often chosen on the basis of pilot studies or the intuition of the original researchers (see Fiedler, 2011, who has outlined the profound effects that such nonrandom samples can have on the outcome of a given study, artificially inflating effect sizes). Each of these choices potentially influences the results of a study and, when ignored, may lead to numerous problems—faulty causal and correlational inferences, initial and replication studies producing divergent results, and studies included in a meta-analysis appearing to come to widely different conclusions (see, e.g., Wainer, 2007; Yong, 2012). In an ideal world, these would not be problems at all. The original researchers would decide and know which sampling decisions are relevant to their research and would provide sufficient description of each. In the real world, however, researchers cannot always know which sampling decisions are important—and, over time, the decisions assumed unimportant may be forgotten.

We take four lessons from these facts. First and obviously, to conduct a direct, systematic, or conceptual replication properly, the replicating researcher needs an

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

adequate description of the original researchers' sampling decisions. Without this description, when the replicating researcher makes inferences regarding the consistency (or reliability) of a finding, he or she cannot determine whether or where his or her sampling decisions differed from the original research. Second, if meta-analysts are to examine quantitatively the robustness of a given phenomenon (e.g., Glass, 1976), they need data permitting them to derive effect sizes (e.g., number of participants and inferential statistics and, more ideally, the central tendency and variance for each measure). Moreover, to evaluate and identify possible moderators of effect sizes, which may reflect sampling decisions in the studies under analysis (e.g., Eysenck, 1994), and to make decisions about how to weight a study,[7] meta-analysts need access to adequate descriptions of the original researchers' sampling decisions. Third, for a reviewer to determine whether an original author met the premises of Mill's (1843) Canons well enough to warrant accurate causal and correlational inferences, an adequate description of sampling decisions must be provided. Finally, if editors and readers alike are to make judgments about the inferential validity of a researcher's conclusions, they need access to descriptions of sampling decisions that determine the appropriate inferences to be made on the basis of a study's outcome. Without access to descriptions of sampling decisions, the process of self-correction in science may be stopped dead in its tracks. In short, the devil lies in the details. These facts lead us back to a duty to describe.

Are the Sampling Decisions of Published Psychological Studies Adequately Described?

As we contemplated these facts, we wondered whether the studies we have published described enough details to permit direct, systematic, or conceptual replication; inclusion in a meta-analysis; or evaluation of our inferential validity. We first reviewed several[8] of the extant inventories designed to provide a "standard way for authors to prepare reports of trial findings, facilitating their complete and transparent reporting, and aiding their critical appraisal and interpretation" (see the Consolidated Standards of Reporting Trials [CONSORT] Statement at http://www.consort-statement.org/). These standards did not suit our purposes. Although each addresses descriptive completeness, none of the demands of Mill's (1843) Canons and accurate inference, replication, or meta-analyses are used in the standards to guide their structure. Moreover, although they include items relevant to these types of studies, other items—such as the formatting details—are not applicable (American Psychological Association, 2009), and some of the standards addressed only intervention designs.

We examined Mill's (1843) Canons and created a comprehensive inventory of the information that a naïve researcher requires to undertake a replication study. We took a deductive approach by first generating a list of items that an original researcher should know, and then we removed items unrelated to Mill's Canons (i.e., researcher hypotheses). Next, we examined the literature and retained only items empirically demonstrated to affect the outcome of studies (for examples from the literature in which they have mattered, see Table S1 in the Supplemental Material available online).[9] If these sampling decisions, known and easily described by the original researcher, have made a difference in some outcomes, then they might also make a difference in other studies. Until the field has a full working knowledge of when and whether they are relevant, it is wise to describe them. For example, few researchers explicitly acknowledge that the time of semester that a study is run might change their sample of student participants, but such an influence has been clearly documented (Aviv, Zelenski, Rallo, & Larsen, 2002; Cassidy & Kangas, 2014).

We modified each item until it appeared directly observable, objective, operationalized, and easily coded; ensured that items were mutually exclusive; and incorporated a scoring metric to guide assessment. The focus of every item is on descriptive sufficiency, and the sole aim is to rate the clarity of description relative to these requirements. We then created a coding manual in which we described each item in enough detail to permit undergraduate research assistants to code the items reliably (see Supplement C in the Supplemental Material).

We applied this inventory to our own work and discovered that we consistently underdescribed sampling decisions made in our experiments and correlational studies. We all too often did not describe variables that could affect the outcome of a study, for example, the sex of the experimenter(s) running our subjects or participants (e.g., Sorge et al., 2014), the exact procedures used (e.g., Bargh et al., 1996), when a study was conducted (e.g., Cassidy & Kangas, 2014), or exactly how we cleaned our data for analysis (for a potent example of the dangers of nondisclosure of data practices, see the exchange between Reinhart & Rogoff, 2010, and Herndon, Ash, & Pollin, 2013). As an example, in one recent case, we could not unearth why our attempts to directly replicate the results of a study performed by a colleague in our laboratory failed because we could not determine how our methodological and training sampling decisions differed from hers—she had not recorded apparently critical decisions, and she could not be reached for a consultation.[10] Moreover, we found that many of the details central to direct replications were not in memory or in the

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

records that remained of our studies. If others requested those details, we simply could not supply them—all potential reasons for a failure to replicate our efforts because of our initial assumption that such details did not matter. Furthermore, without these details available, a replicating researcher may mistakenly conclude the original study was an error.

With that in hand, we asked, "How well are others describing the sampling decisions that support the execution of direct, systematic, or conceptual replication study; inclusion in meta-analysis; accurate inferences; and the analysis of inferential validity of a study?"

Method

The Replicability and Meta-Analytic Suitability Inventory (RAMSI)

The full RAMSI consists of five groups of items known to influence the outcome of a given study that fall within the inferential framework of Mill's (1843) Canons: Methods and Procedures (nine items); Participant Recruitment and Characteristics (nine items); Assessor Recruitment, Characteristics, and Training (13 items); Experimenter Recruitment, Characteristics, and Training (10 items); and Results and Analysis (three items). Each item represents one sampling decision made by the original researcher (for a complete list and description of each of these items, see Supplement A in the Supplemental Material). Because some items apply only to quasi-experimental and experimental designs (e.g., Specific Group Procedures because correlational studies do not contain separate groups), we created two subforms of RAMSI—one for correlational studies and one for experimental and quasi-experimental studies.

In RAMSI, we used a 3-point rating metric for each item (0, 0.5, and 1) and a "not applicable, skip" option when relevant using the following general definitions:

0: Item not found within the study description;

0.5: Item found, but it lacks information necessary for direct replication;

1: Item found, sufficient information to be directly replicable; and

N/A, skip: Item not applicable to study.

The coding manual for RAMSI provided definitions and description for each item (see Supplement C in the Supplemental Material). In the present study, we used a web-based version on the website, DatStat Illume 4.11.[11] When a coder rated an item "N/A" or "0," the web-based inventory skipped all subsequent items related to that topic and assigned an "N/A" or "0" as appropriate. The rating of each item represented how completely the original researcher described the sampling decision that the item addressed.

Articles sampled

We used RAMSI to estimate the average level of relevant descriptive detail of reports published during 2010 in the top five, middle five, and bottom five journals as ranked by the top 100 ISI Web of Science 5-year impact factor. We limited the articles to those describing empirical research, written in English, published in 2010, and involving human participants.[12] We also excluded qualitative studies and meta-analyses. A total of 1,083 articles met inclusion criteria (see Table 2 for a list of journals included and number of articles per journal). We retrieved articles and supplemental materials, when applicable, through the University of Arizona Library resources (e.g., EBSCO, SAGEPub, ISI Web of Science). Articles were assigned unique numbers for coding and tracking purposes.

To calculate reliability scores, two assessors coded about 10% of articles in each journal. An online random number generator (www.random.org) determined the articles to be double-coded. After obtaining the descriptive and reliability data, we tested the hypothesis that higher ranked journals contain more complete research descriptions than lower ranked journals.

Article assessors and data collection

Full details of assessor demographics, recruitment, and procedures, as well as data collection procedures, are available in Supplement D in the Supplemental Material.

Data scoring

Total and section scores, determined as a percentage of points "earned" out of total points possible, were calculated for each article. In other words,

$$\text{Total \% score} = \left(\#\text{ points earned} / \#\text{ points possible}\right) \times 100.$$

All items rated "not applicable" received a point value of 1 because authors made it clear in their report that the item was not applicable and, thus, not an undescribed detail. Subsequently, all items depending on this item were excluded from the sum of total points possible (for specific item point-decision rules, see RAMSI in Supplement A in the Supplemental Material).

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

Table 2. Performance on Replicability and Meta-Analytic Suitability Inventory (RAMSI) Sections and Overall by Journal

RAMSI section	J1 % (*SD*)	J2 % (*SD*)	J3 % (*SD*)	J4 % (*SD*)	J5 % (*SD*)	Rank % (*SD*)
			Top-ranked journals			
Method	49.2 (11.4)	47.4 (10.9)	50.2 (13.4)	45.6 (11.9)	46.7 (12.4)	46.7 (12.0)
Participants	41.7 (23.2)	54.2 (22.4)	69.6 (21.3)	47.7 (20.1)	46.1 (20.0)	50.0 (21.9)
Assessor	70.0 (46.4)	40.0 (44.3)	43.9 (36.9)	45.2 (47.4)	43.8 (47.7)	45.0 (30.3)
Experimenter	25.8 (43.5)	6.7 (21.9)	13.7 (33.8)	10.7 (30.4)	11.5 (30.2)	11.3 (30.3)
Results	68.3 (17.2)	66.5 (21.1)	71.2 (23.7)	68.2 (19.0)	67.6 (23.5)	67.9 (21.3)
Total points	35.4 (13.4)	31.6 (10.9)	39.3 (11.9)	31.3 (13.3)	31.5 (15.6)	32.3 (13.8)
RAMSI section	J6 % (*SD*)	J7 % (*SD*)	J8 % (*SD*)	J9 % (*SD*)	J10 % (*SD*)	Rank % (*SD*)
			Midranked journals			
Method	45.1 (8.4)	40.4 (14.1)	41.6 (8.5)	36.8 (10.2)	39.6 (8.9)	40.9 (11.2)
Participants	52.2 (21.4)	43.0 (24.2)	44.2 (17.6)	32.5 (7.0)	46.3 (19.4)	43.9 (20.9)
Assessor	52.3 (48.2)	42.7 (46.3)	26.4 (37.7)	2.6 (5.6)	42.6 (48.8)	35.7 (44.4)
Experimenter	10.4 (29.8)	24.8 (38.3)	6.6 (24.7)	0.2 (1.0)	10.8 (30.6)	13.6 (31.6)
Results	89.8 (17.0)	51.5 (35.8)	73.8 (28.8)	61.1 (20.7)	59.0 (19.6)	64.9 (31.4)
Total points	35.1 (14.9)	30.9 (18.4)	26.7 (8.0)	18.9 (3.4)	28.9 (16.4)	28.9 (15.1)
RAMSI section	J11 % (*SD*)	J12 % (*SD*)	J13 % (*SD*)	J14 % (*SD*)	J15 % (*SD*)	Rank % (*SD*)
			Low-ranked journals			
Method	43.7 (11.1)	46.6 (9.2)	42.2 (15.7)	43.5 (12.4)	42.0 (14.9)	43.9 (12.3)
Participants	35.7 (8.2)	36.7 (24.2)	62.7 (22.8)	46.3 (20.3)	48.1 (21.4)	45.1 (21.8)
Assessor	36.3 (48.0)	45.1 (47.7)	52.2 (50.4)	47.2 (48.0)	27.8 (42.5)	44.1 (47.4)
Experimenter	11.1 (33.3)	19.4 (36.3)	27.5 (44.3)	12.7 (29.1)	17.4 (33.4)	15.5 (32.4)
Results	50.0 (32.3)	82.1 (26.5)	58.3 (28.9)	71.8 (26.3)	48.0 (29.8)	69.0 (29.1)
Total points	24.0 (3.2)	32.4 (17.2)	35.3 (13.2)	32.0 (16.5)	29.7 (18.5)	31.6 (16.4)

Note: Table 2 illustrates the mean (J# % = journal number mean) and standard deviation (*SD* = journal's standard deviation) percentage of points earned for each section of RAMSI across all articles in each individual journal. These were also calculated across all journals in each journal rank. Mean percentage and standard deviation of total points earned, independent of specific sections assessed, is also provided for articles within each journal. Hence, article total percentage will not equal the average of the percentages earned across the five sections (each consisting of varying number of items), as total points earned divided by total points possible for all articles is averaged in the calculation of "total points." Journals in analyses (J# = number associated with that journal, *n* = number of articles included from that journal) include the following: J1 = *Journal of Experimental Psychology General* (*n* = 40); J2 = *Journal of Personality and Social Psychology* (*n* = 140); J3 = *Development and Psychopathology* (*n* = 55); J4 = *Journal of Cognitive Neuroscience* (*n* = 206); J5 = *Psychological Science* (*n* = 230); J6 = *Personal Relationships* (*n* = 36); J7 = *Journal of Clinical Psychology* (*n* = 79); J8 = *Journal of Personality Assessment* (*n* = 47); J9 = *Applied Psycholinguistics* (*n* = 27); J10 = *Journal of Behavioral Decision Making* (*n* = 26); J11 = *Journal of Creative Behavior* (*n* = 9); J12 = *Teaching of Psychology* (*n* = 40); J13 = *American Journal of Family Therapy* (*n* = 12); J14 = *Psychological Reports* (*n* = 111); J15 = *The Arts in Psychotherapy* (*n* = 25).

Data analyses

We measured assessor reliability as the percentage agreement between assessors. We coded individual items as either *agree* or *disagree* using an "if, then" statement: agreement = 1, and disagreement = 0. Then, we calculated agreement across articles by individual item, section, and total article score using Microsoft Excel.

Each article coded for reliability received two separate sets of ratings. Of the data included in the main analyses, articles were randomly assigned such that the first or second rating was used (50% first rating, 50% second rating). We did this because we had no theoretical reason to think the first rating would be more valid than the second or vice versa. It was also used to avoid inflation of effect sizes by selecting article ratings consistent with hypotheses (e.g., Fiedler, 2011) and to distribute unknown errors.

We tested hypothesized relations among the scores of the three journal groups for the five sections of RAMSI and article total score using one-way, independent-sample analyses of variance (ANOVAs). To detect differences between journal group pairs, when applicable, we used SPSS 20 to conduct post hoc comparisons with Tukey's honestly significant difference and Scheffe's test.

Results

Assessor agreement

Overall, average assessor agreement across all items in the 10% of articles rated by two assessors was 78%. See

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

Table 3. Replicability and Meta-Analytic Suitability Inventory Alternative Brief (RAMSI-AB) Items

Item no.	Item/description
	Method items
1.	Setting: Details regarding setting of study (e.g., dimensions of room, appearance of room, windows)
2.	Study times: Minutes/hours per session, total number of sessions if applicable, time of day and date range of data collection
	Participant items
1.	Recruitment: Description of how participants were recruited
2.	Basic demographics: Age, sex, race/ethnicity, educational status
3.	Inclusion/exclusion/ongoing eligibility criteria: Description of how initial and ongoing eligibility for inclusion in study were determined
	Experimenter/assessor items
1.	Recruitment: Description of how experimenters/assessors were recruited for study
2.	Basic demographics: Age, sex, race/ethnicity, educational status
3.	Training: Description of how experimenters/assessors trained, materials used, any reliability measures taken
4.	Inclusion/exclusion/ongoing eligibility criteria: Description of how initial and ongoing eligibility for inclusion in study were determined

Note: In this table, we list and briefly describe each of the items included in the RAMSI-AB. In the RAMSI-AB, one can mark each item as either in the manuscript, not reported, or not applicable. If unreported, it is suggested that the author explain this reasoning in the cover letter submitted with the manuscript to the editor. The RAMSI-AB represents a subset of important sampling decisions to consider reporting, so authors are advised to reference the RAMSI Alternative as needed and record these when feasible.

Table S2 in the Supplemental Material for item-by-item assessor agreement statistics. It is worth noting that some of the items had low agreement, suggesting that undergraduate students still early in their training find some items particularly difficult to code. Most of the disagreements were between adjacent score points (e.g., assigning an article a .5 vs. 1 or 0 vs. .5) rather than a large discrepancy (e.g., assigning an article a 0 vs. 1). The primary point remains; there is a great deal of description missing from published articles.

Descriptive statistics

Table 3 shows the average percentage scores and standard deviations for each of the journals, by journal rank for the five sections, and total score for RAMSI. There was wide variance of reporting of the sampling decisions surveyed, independent of journal, with 23%–99% either underdescribed (i.e., a score of .5) or undescribed (i.e., a score of 0; for frequency of each score point assigned by item by journal, see Table S3 in the Supplemental Material).

The smallest relative variance was in total scores for Journals 8, 9, and 11. Total scores for Journals 7, 12, and 15 had the largest variance. Examination of score by RAMSI section revealed a wide range of scores both across individual journals and journal rank (for boxplots of article total score by journal, see Figure S1 in the Supplemental Material; for boxplots of score distribution for each RAMSI section by journal rank, see Figure S2 in the Supplemental Material).

Inferential statistics

Using one-way, independent-sample ANOVAs, we compared journal group percentage scores on each section of RAMSI and for total article score (for all ANOVA tables, see Table S4 in the Supplemental Material).

Method. Using an ANOVA, we detected significant differences among journal groups on mean total percentage score for the Method section, $F(2, 1080) = 22.38$, $p < .001$, $r = .20$. Post hoc comparisons indicated that mean score for the top-ranked journal group ($M = 46.7$, $SD = 12.0$) was significantly greater than the midranked ($M = 40.9$, $SD = 11.2$) and the low-ranked ($M = 43.9$, $SD = 12.3$) journal groups. The group mean for the Method section in low-ranked journals was also significantly greater than that of the midranked journal group.

Participants. Using an ANOVA, we detected significant differences among journal groups on Participants total percentage score, $F(2, 1080) = 8.43$, $p < .001$, $r = .12$. Post hoc comparisons indicated that the mean score for the top-ranked journal group ($M = 50.0$, $SD = 21.9$) was significantly greater than the midranked ($M = 43.9$, $SD = 20.9$) and the low-ranked ($M = 45.1$, $SD = 21.8$) journal groups; the mid- and low-ranked journal groups did not differ significantly.

Assessor. Using an ANOVA, we detected significant differences among journal groups on Assessor total percentage score, $F(2, 1080) = 3.42$, $p = .03$, $r = .08$. Post hoc

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

comparisons indicated that the mean score for the top-ranked journal group (M = 45.0, SD = 30.3) was significantly greater than the midranked journal group (M = 35.7, SD = 44.4); the low-ranked journal group (M = 44.1, SD = 47.4) did not differ significantly from either.

Experimenter. There were no detectable differences among journal groups on Experimenter total percentage score, $F(2, 1080) = 1.60$, n.s., $r = .05$ (top-ranked journal group: M = 11.3, SD = 30.3; midranked journal group: M = 13.6, SD = 31.6; low-ranked journal group: M = 15.5, SD = 32.4).

Results. There were no detectable differences among journal groups on Results total percentage score, $F(2, 1080) = 1.62$, n.s., $r = .05$ (top-ranked journal group: M = 67.9, SD = 21.3; midranked journal group: M = 64.9, SD = 31.4; low-ranked journal group: M = 69.0, SD = 29.1).

Article. Using an ANOVA, we detected significant differences among journal groups on Article total score, $F(2, 1080) = 4.48$, $p = .01$, $r = .09$. Post hoc comparisons indicated that mean score for the top-ranked journal group (M = 32.3, SD = 13.8) was significantly greater than the midranked journal group (M = 28.9, SD = 15.1) but not the low-ranked journal group (M = 31.6, SD = 16.4). The mid- and low-ranked groups did not significantly differ.

Discussion

> "The question is not whether various ... effects are real and can be replicated—because they are and often have been—but rather why some researchers reproduce these effects and others do not. The question is important for advancing the knowledge of "how [these] influences operate, and it draws needed attention to the precise contexts and conditions required to produce [them]. More work remains." (Bargh, 2014, p. 36)

Although statistically the top-ranked journals provided more information about methodological and participant sampling decisions than the mid- or low-ranked journals, differences were small. Overall, descriptions were surprisingly incomplete. In the articles in all three journal rankings, the authors most fully described details regarding results (65%–69%) and described the least regarding experimenters (11%–16%). Top-ranked journals provided about 32% of the information regarding an article's overall total sampling decisions, and the middle- and low-ranked journals provided 29% and 32%, respectively. In other words, roughly 70% of the sampling decisions were under- or undescribed, independent of journal rank.

In general, the descriptive completeness of the articles evaluated here appears neither to meet the basic standards set by Mill's (1843) Canons, to allow for full assessment of a researcher's inferential conclusions, nor to meet the needs and requirements of replication or meta-analytic researchers. Moreover, although some statistically significant differences among reporting in top-, mid-, and low-ranked psychological journals appeared, the differences were relatively small; the lack of descriptive completeness across the journals sampled provides little substantial evidence of higher editorial standards by journal rank.

There is no simple fix for the problem at hand (see the Cochrane Review by Turner, Shamseer, Altman, Schulz, & Moher, 2012, on the effects of the CONSORT Statement on completeness of reporting in the medical field and the need for both author and journal involvement). McShane and Böckentholt (2014, this issue), for example, have called attention to the fact that the heterogeneity of effect sizes across replication attempts may lead to the mistaken interpretation that a replication effort failed. These authors have advised replicating researchers to gain estimates of heterogeneity of effect sizes (e.g., by having several labs simultaneously run the same study worldwide—the Many Labs approach; Klein et al., 2014) and then to use adjusted power analyses to guide the N in their participant sample. In short, whereas McShane and Böckentholt (2014) have called for researchers to adjust for heterogeneity of effect sizes because of unknown error, our call is for researchers to provide information that would allow for the identification of sources of that heterogeneity. Taken together, McShane and Böckentholt have provided a means of dealing with heterogeneity potentially attributable to under- or undescribed sampling decisions, and we provide a means of minimizing such heterogeneity. In our opinion, combining these approaches strengthens both.

We encourage researchers to use checklists, such as the RAMSI (or one of the alternative versions presented later), to guide the descriptive content of their reports. As much as comprehensive checklists have been shown to reduce complications in hospital operating rooms (de Vries et al., 2010), our inventory has the potential to reduce descriptive incompleteness in psychology.

Potential applications of the RAMSI

We offer three versions of RAMSI that researchers and editors, among others, may find useful in the Supplemental Material. The first (see Supplement A in the Supplemental Material) is the original version of RAMSI that we used and described in this article. It was created so that the current literature could be evaluated.

The second version, the RAMSI Alternative (RAMSI-A; see Supplement E in the Supplemental Material), could

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

be helpful for both publication and general use. It guides users through each RAMSI item, provides the option of marking an item as already described in the manuscript (if used for publication) or as "N/A" if an item is not applicable to a given study, and contains space to provide a description of the sampling decision. It can be used by researchers to record crucial sampling decisions when conducting a study (for a demonstration of its use as applied to this article, see Supplement D in the Supplemental Material).

The RAMSI Alternative Brief (RAMSI-AB), created in light of the data from the present study, is a nine-item form notably shorter than the 44-item RAMSI and RAMSI-A (see Table 3). We omitted items that authors routinely and fully describe, that are obvious to report (inferential statistics, method of analysis, procedures used, etc.), and that are covered in other recent recommendations (as noted later in this section). Several related individual items were collapsed into single items because there were not substantial differences in level of published descriptions among them (e.g., assessor training vs. assessor reliability). Furthermore, because many psychological studies do not use assessors, assessor and experimenter items were placed into a single category. Thus, RAMSI-AB includes the core sampling decisions that are typically most under- or undescribed as informed by the data, literature, and Mill's (1843) Canons. The RAMSI-AB is one possible and easily implemented set of recommendations regarding description of sampling decision that extends beyond recent recommendations about methodological disclosure alone.

Likely, the RAMS-AB would be most useful when dealing with a specific manuscript to track which items were included, not applicable, or omitted, and it could serve as a quick overview of such decisions to editors and reviewers alike. When items are not reported, we make the additional recommendation that the author(s) explain the reasoning behind this omission in the cover letter submitted with the manuscript for publication. Authors would then be on record about which sampling decisions replicating studies would be free to vary as desired.

We recommend that researchers use checklist items to record the sampling decisions they make when designing a study and before running it. We further recommend that after completing a study, researchers make conscious, deliberate, and intentional decisions—informed by Mill's (1843) Canons and what is known in the literature—to determine which items should be described to help foster inferential validity, replicability, meta-analytic inclusion, and, thereby, self-correcting science. Finally, though we acknowledge the possibility that some details might prove to be unimportant, we recommend that researchers describe items reported in publications or online supplemental materials in enough detail to warrant direct replication by independent researchers.

Objections and rebuttals

Some may argue that the problem identified here is trivial and may assert that this article is unnecessary because a simple solution is already at hand. A reviewer (received April 26, 2012) of an earlier version of this article dismissed these ideas by commenting,

> I do not believe that the barrier to conducting these replications lies in the inability of researchers to attain original materials or sufficient details about the published study that they are attempting to replicate. Except in unusual cases, it is not difficult to contact authors to request their exact materials.

The reviewer's belief is correct. It is not difficult to request exact materials. Furthermore, with the exception of authors who are no longer in the field, retired, or dead, the problem does not lie in making contact. Instead, the problem lies in the authors' responses to those contacts. There are more than 50 years of data demonstrating that most authors simply do not respond to requests for methodological details, exact materials, or raw data. Many who do respond cannot supply requested information because the information was never recorded in the first place, has been forgotten, or was lost. In his undergraduate classes at Harvard, Robert Rosenthal vividly illustrated the latter by stating there are "an unusual number of fires and floods in Psychology Departments" (Kevin Thomas, personal communication, December 2012).

To our knowledge, Wolins (1962) was the first to document this problem. Of the 37 authors contacted for further information, only 24% provided it without insisting on having control of subsequent publications. The introduction of e-mail has not fixed the situation. Wicherts and colleagues (Wicherts, Bakker, & Molenaar, 2011; Wicherts, Borsboom, Kats, & Molenaar, 2006) requested study information from contemporary authors and found that 42.9% provided the requested information. More recently, LeBel et al. (2013) requested additional methodological descriptions from authors published in four major psychology publications;[13] they received an overall response rate of 46.4%. Others have documented widespread failures to adhere to ethical guidelines and journal policies for disclosure at time of publication (e.g., Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2011).[14] It appears that the reviewer's belief is correct, but the implication that contact solves the problem is not.

Others might argue that journals are already implementing standards addressing this issue. The editorial board members of *Nature* were among the first

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

to recognize the problems that incomplete descriptions create. In an April 2013 editorial, it was recognized that, although problems of replication start in the laboratories of researchers, journals contribute "when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly" (Nature Publishing Group, 2013, p. 398). As of May 2013, the editorial board of *Nature* implemented editorial practices "to ease the interpretation and improve the reliability of published results" by giving more space to method sections and ensuring that authors describe "key methodological details" through use of a checklist (Nature Publishing Group, 2013, p. 398). A similar series of policy changes has been recently implemented in *Psychological Science* (influenced by LeBel et al., 2013), requiring authors to confirm reporting of observation exclusions, all manipulations, all measures analyzed, and sample size decisions in articles (Eich, 2014).[15]

These changes are important but incomplete. They are neither anchored in the requirements of Mill's (1843) Canons nor focused on ensuring the inferential validity of a study, its replicability, or its inclusion in meta-analyses. Thus, they do not adequately foster self-correcting science. They also do not fully take into account what is known from the literature regarding sampling decisions that may influence a study's outcome (e.g., sex of the experimenter, exact procedures used, when a study was conducted). Moreover, the data collected here demonstrate that far more than the handful of methodological sampling decisions tagged as needing to be described by past researchers are underreported or undescribed. Most glaring, this list includes crucial description of 80%–99% of sampling decisions involving experimenters and assessors, study settings, and time parameters of a study. Less glaring, though still troublesome, the studies in our sample also left underreported or undescribed 60%–90% sampling decisions relevant to participant demographics, recruitment, and participation eligibility criteria. These data suggest that although authors describe what they think matters, crucial sampling decisions are ignored, and authors, editors, or readers cannot rely on intuition alone to determine when and whether such details are important to record. All of these omissions may inhibit not only replication and meta-analysis but also any causal and correlational inferences based on research findings. Given that these sampling decisions are taken into account with RAMSI, RAMSI-A, and RAMSI-AB, they are more comprehensive than past suggestions.

Some also might argue that the level of descriptive detail called for by Mill's (1843) Canons and embodied in RAMSI is too stringent because publishing space comes at too significant a cost, making this level of detail unrealistic, and researchers cannot be expected to describe so much about their studies.

The technological advances of the past 20 years, however, afford us inexpensive opportunities to post near unlimited supplemental material online. Even adherence to the considerably briefer RAMSI-AB would be an improvement on current standards, encouraging authors and editors to consider reporting decisions actively. Moreover, placing information, such as that contained in RAMSI-A, online (ideally with online supplemental material maintained by the publishing journal to ensure that it does not disappear and that it remains linked to the published study) could save replicating researchers time, effort, and expense. In some cases, it could also save original researchers embarrassment as well as save politicians from advocating or establishing harmful public policy (see again the exchange between Reinhart & Rogoff, 2010, and Herndon et al., 2013). In the present analysis, multiple articles merited the directly replicable score of "1" for each of the inventory items, indicating that this level of description can and already is achieved by at least some authors in some journals. Furthermore, all RAMSI-A items are easily known to and recordable by an original researcher—problems arise when the researcher erroneously assumes that an item is unimportant and fails to report it. In short, the argument that it is too expensive or too difficult to provide full descriptions of pertinent sampling decisions has lost its weight in today's technological climate, especially in light of the damage that not providing them creates.

Finally, some might object that the different versions of RAMSI are incomplete. The items are limited to variables demonstrably relevant, at least in some cases, to the outcome of a study and to those sampled by, known to, easily recorded by, and in the control of a researcher. Researchers in the field do not yet have a good understanding of how to predict whether or when or even which of these sampling decisions will be relevant to the outcome of a given study.[16] Despite the incompleteness of RAMSI-AB and even RAMSI-A, implementation of such standards of reporting of sampling decisions represents a step toward a solution to the problem by helping researchers in the field build a base knowledge, better allowing examination of when and whether these sampling decisions even matter.

Final thoughts

We offer these thoughts and data in the hope that many of the problems facing a self-correcting science founded in valid inferences will continue to receive careful and thoughtful attention and intervention. Moreover, we hope that researchers in the field will keep an eye on the logic underpinning their designs, strategies used by other areas of science, and journals such as *Nature* and *Psychological Science* as more rigorous publication guidelines are

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

instituted. No doubt there will be many strong and well-thought-out options forthcoming. In our opinion, researchers in the field would benefit by basing such options on intentional, principled, and conscious choice informed by research and foundational principles expressed by John Stuart Mill and others rather than by leaving the description of sampling decisions to chance, uniformed choice, or intuition alone.

Acknowledgments

We acknowledge and thank Alison Ledgerwood for her hard work, patience, and suggestions during the development of this article. We also thank the many undergraduate research assistants who contributed to this project.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information may be found at http://pps.sagepub.com/content/by/supplemental-data

Notes

1. Direct replication is an attempt to reproduce a study's finding with the same suite of samples, from participants to procedures, as the original study.
2. Systematic replication is an attempt to reproduce a study's findings with the same general methods as the original study but with modifications in some, but not all, of the sampling decisions.
3. Conceptual replication is an attempt to reproduce a study's finding with samples from conceptually related population spaces.
4. See *Science* (December 2011) and *Perspectives in Psychological Science* (November 2012 and May 2014).
5. We use the word "decision" as a convenience, recognizing that these sampling decisions are often imposed. For example, a researcher may not have a choice regarding laboratory space or sex of research assistants.
6. It is worth mentioning that whether these replication efforts were successful remains a matter of intense debate.
7. This applies if the meta-analytic researcher intends to take method quality or even participant sample size into account.
8. We reviewed the Consolidated Standards of Reporting Trials (CONSORT; see http://www.consort-statement.org), the Journal of Article Reporting Standards (JARS), and the Meta-Analysis Reporting Standards (MARS; e.g., see http://www.apastyle.org/manual/related/JARS-MARS.pdf); the Transparent Reporting of Evaluation of Nonrandomized Designs Statement and the Preferred Items for Systematic Reviews of Meta-Analyses (Begg et al., 1996; Des Jarlais, Lyles, Crepaz, & the TREND Group, 2004; Moher, Liberati, Tetzlaff, & Altman, 2009); and the *Publication Manual of the American Psychological Association* (6th ed.; American Psychological Association, 2009).
9. An exhaustive review of literature pertaining to each item is far beyond the scope of this article—Table S1 in the Supplemental Material should be viewed as a (nonrandom) sample of the literature rather than representing its entirety.
10. We should also note that three of the authors have several studies that failed to replicate within their own labs and so were, with one exception, never published. In light of the potential importance of knowing the sampling decisions made, we now dearly wish that all of these decisions had been recorded so we might generate hypotheses designed to explain why we failed to replicate our own studies convincingly.
11. Unfortunately, the University of Arizona terminated its contract with DatStat Illume, so the original online version of the RAMS Inventory is no longer available; however, the paper version can be found in Supplement A in the Supplemental Material.
12. We examined only studies in which human participants were used, but the principles discussed in this article are also applicable to nonhuman subject research, as Sorge et al. (2014) have demonstrated.
13. The four major psychology publications include the following: *Psychological Science; Journal of Personality and Social Psychology; Journal of Experimental Psychology: Learning, Memory, and Cognition*; and *Journal of Experimental Psychology: General*.
14. The problem of incomplete description is not a problem of psychology alone (for further examples from other fields, see Reidpath & Allotey, 2001; Savage & Vickers, 2009). We may be particularly sensitive about this point. While conducting a meta-analysis examining the efficacy of cognitive, behavioral, and cognitive behavioral interventions, we have had similar difficulties getting authors to respond to requests for information about their methods, data, and statistics. Our conceptual replication of sorts got about the same pattern of results as Wicherts and colleagues, forcing us to exclude a number of articles from our analysis.
15. See the 2014 submission guidelines for *Psychological Science* at http://www.psychologicalscience.org/index.php/publications/journals/psychological_science/ps-submissions.
16. A good example of a recent study (in which one of the authors of this article participated) in which this issue was tackled is the "Many Labs" study by Klein et al. (2014). The authors looked at 13 published psychological findings to see how well they would replicate with diverse participant samples in different settings. On the whole, the results were encouraging, with 10 of the 13 effects replicating across participant samples (U.S. vs non-U.S.) or settings (online vs. in lab). At the same time, the sample and setting variables were significant moderators for about one third of the effects reported.

References

*Indicates works cited in Table S1

*Alcock, N. (1996). Factors affecting the assessment of post-operative pain: A literature review. *Journal of Advanced Nursing*, *24*, 1144–1151.

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public availability of published research data in high-impact journals. *PLoS ONE*, *6*(9), e24357. doi:10.1371/journal.pone.0024357

American Psychological Association. (2009). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119. doi:10.1002/per.1919

Aviv, A. L., Zelenski, J. M., Rallo, L., & Larsen, R. J. (2002). Who comes when: Personality differences in early and late participation in a university subject pool. *Personality and Individual Differences*, *33*, 487–496.

*Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training session on rate of learning to type. *Ergonomics*, *21*, 627–635.

*Baker, M. D., & Maner, J. K. (2009). Male risk-taking as a context-sensitive signaling device. *Journal of Experimental Social Psychology*, *45*, 1136–1139.

Bargh, J. A. (2014). Our unconscious mind: Unconscious impulses and desires impel what we think and do in ways Freud never dreamed of. *Scientific American*, *310*, 30–37.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., . . . Stroup, D. F. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *Journal of the American Medical Association*, *276*, 637–639.

*Bernstein, L. (1956). The examiner as inhibiting factor in clinical testing. *Journal of Consulting Psychology*, *20*(4), 287–290.

*Bierenbaum, H., Nichols, M. P., & Schwartz, A. J. (1976). Effects of varying session length and frequency in brief emotive psychotherapy. *Journal of Consulting and Clinical Psychology*, *44*, 790–798.

Bordens, K. S., & Abbott, B. B. (2008). *Research design and methods: A process approach* (7th ed.). Boston, MA: McGraw Hill.

*Campbell, T. S., Holder, M. D., & France, C. R. (2006). The effects of experimenter status and cardiovascular reactivity on pain reports. *Pain*, *125*, 264–269.

Cassidy, R. N., & Kangas, B. D. (2014). Impulsive students articulate later: Delay discounting in a research subject pool. *The Experimental Analysis of Human Behavior Bulletin*, *30*, 1–5.

Cesario, J., Plaks, J. E., & Higgins, E. T. (2006). Automatic social behavior as motivated preparation to interact. *Journal of Personality and Social Psychology*, *90*, 893–910. doi:10.1037/0022-3514.90.6.893

*Chapman, L. J., & Chapman, J. P. (1978). The measurement of differential deficit. *Journal of Psychiatric Research*, *14*, 303–311.

*Cohen, D., Atun-Einy, O., & Scher, A. (2012). Seasonal effects on infants' sleep regulation: A preliminary study of Mediterranean climate. *Chronobiology International*, *29*, 1352–1357.

Cohen, M. R., & Nagel, E. (1934). *An introduction to logic and scientific method*. New York, NY: Harcourt, Brace.

*Cooper, H., Baumgardner, A. H., & Strathman, A. (1991). Do students with different characteristics take part in psychology experiments at different times of the semester? *Journal of Personality*, *59*, 109–127. doi:10.1111/j.1467-6494.1991.tb00770.x

*Curtis, H. S., & Wolf, E. (1951). The influence of the sex of the examiner on the prediction of sex responses on the Roschach. *American Psychologist*, *6*, 345–346.

*Davis, D. W. (1997). The direction of race of interviewer effects among African-Americans: Donning the black mask. *American Journal of Political Science*, *41*, 309–322.

Des Jarlais, D. C., Lyles, C., & Crepaz, N., & the TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The trend statement. *American Journal of Public Health*, *94*, 361–366.

*deTurck, M. A., & Miller, G. R. (1990). Training observers to detect deception: Effects of self-monitoring and rehearsal. *Human Communication Research*, *16*, 603–620.

de Vries, E. N., Prins, H. A., Crolla, R. M., den Outer, A. J., van Andel, G., & van Helden, S. H., . . . SURPASS Collaborative Group (2010). Effect of a comprehensive surgical safety system on patient outcomes. *New England Journal of Medicine*, *363*, 1928–1937. doi:10.1056/NEJMsa0911535

Doyen, S., Klein, O., Pinchon, C., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*(1), e29081. doi:10.1371/journal.pone.0029081

*Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A., Cronin, E., . . . Williamson, P. R. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS Clinical Trials*, *3*(8), e3081. doi:10.1371/journal.pone.0003081

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*(3), 3–6. doi:10.1177/0956797613512465

*Epley, N., & Huff, C. (1998). Suspicion, affective response, and educational benefit as a result of deception in psychology research. *Personality and Social Psychology Bulletin*, *24*, 759–768.

Eysenck, H. J. (1994). Meta-analysis and its problems. *British Medical Journal*, *309*, 789–792.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for controlling and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128.

Fiedler, K. (2011). Voodoo correlations are everywhere—Not only in neuroscience. *Perspectives on Psychological Science*, *6*, 163–171. doi:10.1177/1745691611400237

*Fucci, D., Petrosino, L., Sloane, N., & Cantrell, J. (1981). Source variation on lingual vibrotactile thresholds: I. The influence of experimenter experience. *Bulletin of the Psychonomic Society*, *17*, 231–232.

*Gamboz, N., Russo, R., & Fox, E. (2002). Age differences and the identity negative priming effect: An updated meta-analysis. *Psychology and Aging*, *17*, 525–530.

*Garb, H. (1998). *Studying the clinician*. Washington, DC: American Psychological Association.

*Giacomoni, C., & Davies, P. (2013). Effect of room characteristics on perception of low-amplitude sonic booms heard indoors. *Proceedings of Meetings on Acoustic*, *19*, 040050. Retrieved from http://asadl.org/poma/resource/1/pmarcw/v19/i1/p040050_s1

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *The Educational Researcher*, *10*, 3–8.

Goodwin, C. J., & Goodwin, K. A. (2013). *Research in psychology: Methods and design* (7th ed.). Hoboken, NJ: John Wiley & Sons.

Grant, S., Mayo-Wilson, E., Hopewell, S., Macdonald, G., Hoher, D., & Montgomery, P. (2013). Developing a reporting guideline for social and psychological intervention trials. *Journal of Experimental Criminology*, *9*, 355–367. doi:10.1007/s11292-013-9180-5

*Harris, S. (1971). Influence of subject and experimenter sex in psychological research. *Journal of Consulting and Clinical Psychology*, *37*, 291–294.

Herndon, T., Ash, M., & Pollin, R. (2013). *Does high public debt consistently stifle economic growth?: A critique of Reinhart and Rogoff* (Working paper series, University of Massachusetts at Amherst. Political Economy Research Institute, No. 322). Retrieved from http://www.peri.umass.edu/fileadmin/pdf/working_papers/working_papers_301-350/WP322.pdf

*Hsieh, A. Y., Tripp, D. A., & Ji, L. (2011). The influence of ethnic concordance and discordance on verbal reports and nonverbal behaviors of pain. *Pain*, *152*, 2016–2022.

Hull, J., Slone, L., Metayer, K., & Matthews, A. (2002). The nonconsciousness of self-consciousness. *Journal of Personality and Social Psychology*, *83*, 406–424.

*Imeraj, L., Antrop, I., Roeyers, H., Swanson, J., Deschepper, E., Bal, S., & Deboutte, D. (2012). Time-of-day effects in arousal: Disrupted diurnal cortisol profiles in children with ADHD. *Journal of Child Psychology and Psychiatry*, *53*, 782–789. doi:10.1111/j.1469-7610.2012.0256.x

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

Ioannidis, J. P. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice*, *14*, 951–957. doi:10.1111/j.1365-2753.2008.00986.x

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

*Jacobs, W. J., Blackburn, J. R., Butrick, M., Harpur, T. J., Kennedy, D., Mana, M. J., MacDonald, M. A., . . . Pfaus, J. G. (1988). Observations. *Psychobiology*, *16*, 3–19.

*Kallai, I., Barke, A., & Voss, U. (2004). The effects of experimenter characteristics on pain reports in women and men. *Pain*, *112*, 142–147.

*Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010). The impact of outcome reporting bias in randomized controlled trials on a cohort of systematic reviews. *British Medical Journal*, *340*, c365. doi:10.1136/bmj.c365

*Klainin-Yobas, P., Cho, M. A. A., & Creedy, D. (2012). Efficacy of mindfulness-based interventions on depressive symptoms among people with mental disorders: A meta-analysis. *International Journal of Nursing Studies*, *49*, 109–121.

Klatzky, R. L., & Creswell, J. D. (2014). An intersensory interaction account of priming effects—and their absence. *Perspectives on Psychological Science*, *9*, 48–58. doi:10.11/1456916135144468

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152. doi:10.1027/a000001

LeBel, E. P., Borsbook, D., Giner-Sorolla, R., Hasselman, F., Peters, K., Ratliff, K. A., & Smith, C. T. (2013). PsychoDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science*, *8*, 424–432. doi:10.1177/1745691613491437

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*, 371–379. doi:10.1037/a002525172

*Lew, M. B. (1982). Child and adult experimenters: Some differential effects. *Child Study Journal*, *12*(4), 223–235.

Lilienfeld, S. O., Lynn, S. J., Namy, L., & Woolf, N. (2009). *Psychology: From inquiry to understanding*. Boston, MA: Allyn & Bacon.

*Little, L. M., Curran, J. P., & Gilbert, F. S. (1977). The importance of subject recruitment procedures in therapy analogue studies on heterosexual-social anxiety. *Behavior Therapy*, *8*, 24–29.

*Lombardo, J. P., & Tocci, M. E. (1979). Attribution of positive and negative characteristics of instructors as a function of attractiveness and sex of instructor and sex of subject. *Perceptual & Motor Skills*, *48*, 491–494. doi:10.2466/pms.1979.48.2.491

*Lowrey, P. E. (1993). The Assessment Center: An Examination of the Effects of Assessor Characteristics on Assessor Scores. *Public Personnel Management*, *22*, 487–501.

*Macoby, E. (1974). *The psychology of sex differences vol. I: Text*. Stanford, CA: Stanford University Press.

*Matsumoto, D. (1993). Ethnic differences in affect intensity, emotional judgments, display rule attitudes, and self-reported emotional expression in an American sample. *Motivation and Emotion*, *17*, 107–123.

*May, C. P., Hasher, L., & Foong, N. (2005). Implicit memory, age, and time of day: Paradoxical priming effects. *Psychological Science*, *16*, 96–100. doi:10.1111/j.0956-7976.2005.00788.x

McShane, B. B., & Böckentholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives in Psychological Science*, *9*, 612–625.

Mill, J. S. (1843). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation* (Vols. 1–3). London, England: John W. Parker, West Strand.

*Mintzer, E., & Halpern, J. (1980). Effect of sex of therapist and client on therapists' attitudes toward assertiveness problems. *Journal of Clinical Psychology*, *36*, 704–708. doi:10.1002/1097-4679(198007)36:3<704::AID-JCLP2270360317>3.0.CO;2-F

*Mishra, S. P. (1980). The influence of examiners' ethnic attributes on intelligence test scores. *Psychology in the Schools*, *17*, 117–122.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Medicine*, *6*(6), e1000097. doi:10.1371/journal.pmed1000097

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

Nature Publishing Group. (2013). Announcement: Reducing our irreproducibility. *Nature*, *496*, 398. doi:10.1038/496398a

Pashler, H., Harris, C., & Coburn, N. (2011, September 15). *Elderly-related words prime slow walking*. Retrieved from http://www.PsychFileDrawer.org/replication.php?attempt=MTU%3D

*Paulus, P. B., Annis, A. B., Seta, J. J., Schkade, J. K., & Matthews, R. W. (1976). Density does affect task performance. *Journal of Personality and Social Psychology*, *34*, 248–253. doi:10.1037/0022-3514.34.2.248

*Pereira, M., & Austrin, H. R. (1980). Locus of control and status of the experimenter as predictors of suggestibility. *Clinical and Experimental Hypnosis*, *28*, 367–374. doi:10.1080/00207148008409865

Petri, A. (2014, April 28). The best-laid plans of lab mice and men: Are male researchers ruining mouse science? *The Washington Post*. Retrieved from http://www.washingtonpost.com/blogs/compost/wp/2014/04/28/the-best-laid-plans-of-lab-mice-and-men-are-male-researchers-ruining-mouse-science/

Reidpath, D. D., & Allotey, P. A. (2001). Data sharing in medical research: An empirical investigation. *Bioethics*, *15*, 125–134. doi:10.3758/BF03194105

Reinhart, C. M., & Rogoff, K. S. (2010). Growth in a time of debt (NBER Working Paper Series). *American Economic Review, American Economic Association*, 100, 573–578.

*Rodger, R. S., & Roberts, M. (2013). Comparison of power for multiple comparison procedures. *Journal of Methods and Measurement in the Social Science*, *4*(1), 20–47.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*(2). Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2012/february-12/psychologys-woes-and-a-partial-cure-the-value-of-replication.html

*Roll, J. M., McSweeney, F. K., Cannon, C. B., & Johnson, K. S. (1996). Knowledge of session length is determinant of within-session response patterns in human operant paradigm. *Behavioural Processes*, *36*, 1–9.

*Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York, NY: Appleton-Century-Crofts.

*Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject*. New York, NY: Wiley.

*Rüger, M., Gordijn, M. C., Beersma, D. G., de Vries, B., & Daan, S. (2006). Time-of-day-dependent effects of bright light exposure on human psychophysiology: Comparison of daytime and nighttime exposure. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology*, *290*, R1413–R1420. doi:10.1152/ajpregu.00121.2005

*Rumenik, D. K., Capasso, D. R., & Hendrick, C. (1977). Experimenter sex effects in behavioral research. *Psychological Bulletin*, *84*, 852–877. doi:10.1037/0033-2909.84.5.852

*Saunders, D. R. (1980). Definition of Stroop interference in volunteers and non-volunteers. *Perceptual & Motor Skills*, *51*, 343–354.

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE*, *4*(9), e7078. doi:10.1371/journal.pone.0007078

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100. doi:10.1037/a0015108

*Schmitt, N., Schneider, J. R., & Cohen, S. A. (1990). Factors affecting validity of a regionally administered assessment center. *Personnel Psychology*, *43*, 1–12.

Shadish, W. R., Doherty, M., & Montgomery, L. M. (1989). How many studies are in the file drawer? An estimate from family/marital psychotherapy literature. *Clinical Psychology Review*, *9*, 589–603. doi:10.1016/0272-7358(89)90013-5

*Sholomskas, D. E., & Syracuse-Siewert, G. (2005). We don't train in vain: A dissemination trial of three strategies of training clinicians in cognitive-behavioral therapy. *Journal of Consulting and Clinical Psychology*, *73*, 106–115. doi:10.1037/0022-006X.73.1.106

Sidman, M. (1960). *Tactics of scientific research*. New York, NY: Basic Books.

*Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760.

Sorge, R. E., Martin, L. J., Isbester, K. A., Sotocinal, S. G., Rosen, S., Tuttle, A. H., . . . Mogil, J. S. (2014). Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nature Methods*. Advance online publication. doi:10.1038/NMETH.2935

*Steer, R. A., Beck, A. T., Riskind, J. H., & Brown, G. (1987). Relationships between the Beck Depression Inventory and the Hamilton Psychiatric Rating Scale for Depression in depressed outpatients. *Journal of Psychopathology and Behavioral Assessment*, *9*, 327–339.

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

*Tanke, E. D. (1979). Perceptions of ethicality of psychological research: Effects of experimenter status, experiment outcome, and authoritarianism. *Personality and Social Psychology Bulletin*, *5*, 164–168.

Turner, L., Shamseer, L., Altman, D. G., Schulz, K. F., & Moher, D. (2012). Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Systematic Reviews*, *1*, Article 60. doi:10.1186/2046-4053-1-60

*van der Heijden, K. B., & de Sonneville, L. M. J. (2010). Time-of-day effects on cognition in preadolescents: A trails study. *Chronobiology International*, *27*, 1870–1894. doi:10.3109/07420528.2010.516047

*Vieluf, S., Godde, B., Reuter, E., & Voelcker-Rehage, C. (2013). Age-related differences in finger force control are characterized by reduced force production. *Experimental Brain Research*, *224*, 107–117. doi:10.1007/s00221-012-3292-4

*Voyer, D., Voyer, S., & Bryden, B. M. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250–270.

Wainer, H. (2007). The most dangerous equation: Ignorance of how sample size affects statistical variation has created

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016

havoc for nearly a millennium. *American Scientist*, *95*(3), 249–256. doi:10.1511/2007.65.1026

*Weiss, B. H., O'Mahoney, M., & Wichchukit, S. (2010). Various paired preference tests: Experimenter effect on take home choice. *Journal of Sensory Studies*, *25*, 778–790.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to strength of the evidence and the quality of reporting statistical results. *PLoS ONE*, *6*(11), e26828. doi:10.1371/journal.pone.0026828

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.

*Wilkins, W. (1986). Therapy-therapist confounds in psychotherapy research. *Cognitive Therapy and Research*, *10*, 3–11.

*Williams, D. A., & Thorn, B. E. (1986). Can research methodology affect treatment outcome? A comparison of two cold pressor test paradigms. *Cognitive Therapy and Research*, *10*, 539–545.

*Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management*, *29*, 231–258.

*Wolchik, S. A., Spencer, S. L., & Lisi, I. S. (1983). Volunteer bias in research employing vaginal measures of sexual arousal. *Archives of Sexual Behavior*, *12*, 399–408.

Wolins, L. (1962). Responsibility for raw data. *American Psychologist*, *17*, 657–658.

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*, 298–300. Retrieved from http://www.nature.com/news/replication-studies-bad-copy-1.10634

*Yoon, C., Lee, M. P., & Danziger, S. (2007). The effects of optimal time of day on persuasion processes in older adults. *Psychology & Marketing*, *24*, 475–495. doi:10.1002/mar.20169

*Zagar, R., & Bowers, N. D. (1983). The effect of time of day on problem solving and classroom behavior. *Psychology in Schools*, *20*, 337–345.

Downloaded from pps.sagepub.com at PENNSYLVANIA STATE UNIV on May 17, 2016