

## A General Framework for Curating Replication Evidence of Social Science Findings

Every year, society spends billions of dollars (primarily of public tax payer money) to fund scientific studies aimed at deepening our understanding of the natural and social world. The hope is that the findings yielded by these studies will help us address important societal problems (e.g., cancer; suicide; racial discrimination; voter turnout). The findings yielded by these studies, however, can only be considered sufficiently trustworthy knowledge ready to inform public-policy decisions once they have been successfully *replicated* and *generalized* by independent researchers. *Successful replication* is taken to mean that independent researchers have been able to consistently observe similar results as originally reported using *similar* methodology and conditions to an original study. *Successful generalization* is taken to mean that independent researchers have been able to consistently observe similar results as originally reported under situations that use *different* methodologies (often superior methodologies or measurement instruments), contexts, and populations, consequently producing evidence that original results *generalize* to these different situations.<sup>1</sup>

To achieve our goal of creating trustworthy knowledge then, we need to systematically track the *replicability* and *generalizability* of social science findings over time. The following article proposes a general and unified framework for the tracking and curation of replicability and generalizability evidence of social science findings, with the goal of producing a dynamic, living, and continuously evolving body of **knowledge** that can soundly inform public-policy. The general framework needs to be very flexible to overcome several distinct conceptual, epistemological, and statistical challenges that arise when tracking and gauging replicability and generalizability. Each of the following challenges needs to be overcome to achieve our goal:

1. Accommodation of different approaches to replication: The current focus in economics and political science is on **analytic reproducibility** and **robustness analyses** whereas current focus in psychology is on **new sample replications**. Each of these approaches is important and the order in which these approaches is implemented is crucial to maximize research efficiency. Findings that are not analytically reproducible and/or analytically robust may not be worth the costly expenses required to attempt to replicate in a new sample. Also, for maximal knowledge creation, it is crucial to have interdisciplinary curation of replication evidence rather than having economists, political

---

<sup>1</sup> Current approaches (i.e., traditional meta-analyses ) to synthesizing evidence are unable to produce the trustworthy knowledge we seek because these cannot fully account for publication bias (Ferguson & Heene, 2012; McShane, Bockenholt, & Hansen, 2016; Rosenthal, 1979), questionable research practices (John et al., 2012), and unintentional exploitation of design and analytic flexibility (Simmons et al., 2011; Gelman & Loken, 2013) and the various unknowable interactions among these factors.

scientists, and psychologists maintain their own replication databases (as is currently the case).

2. Accumulation of replication evidence that speaks to the *replicability* and *generalizability* of an effect/hypothesis (i.e., *Replicability* replication evidence vs. *Generalizability* replication evidence). To achieve this, we need flexible ontological structures – what we’re calling *evidence collections* – to accommodate replication studies of specific effects/hypotheses being nested in different ways in relation to original studies that test an effect across different generalizations and/or operationalizations of the target constructs (e.g., *replications of an effect via a single vs. multiple generalization(s)/operationalization(s) originating from a single published article; replications of an effect via multiple generalizations and/or operationalizations originating from several different published articles*).
3. Accommodation of different kinds of studies (e.g., experimental studies [RCTs], observational and correlational studies) and study designs (e.g., between-subjects designs, within-subject designs, interaction designs, etc.).
4. Development of a working replication taxonomy to allow a principled and justifiable approach to distinguishing replications that are sufficiently methodologically similar to an original study vs. insufficiently methodologically similar. Such a taxonomy also guides what kind of original studies are eligible to be included in evidence collections as separate generalization branches under which direct replication studies are curated.
5. Taking into account study quality of replications and ability to pool across different subsets of replications that vary on the following study quality dimensions: (1) verifiability (e.g., open data/materials availability), (2) pre-registration status, (3) analytic reproducibility verification status, (4) analytic robustness verification status, (5) active sample evidence (also known as *positive controls*), and (6) replication design differences.
6. Development of a principled approach to meta-analytically combining replication evidence within and across generalizations of an empirical effect and interpreting the overall meta-analytic results (e.g., fixed-effect vs. random-effects model, possibly hierarchical in the case of multiple generalizations and correlated outcomes; Bayesian approaches to yield more principled and meaningful credible intervals; small telescope approach in the case of very few replication studies).
7. Creation of a viable crowd-sourcing system that includes key features to (i) incentivize number and frequency of contributions (low-barrier-to-entry approach, user contributions

prominently displayed on public user profile and home page) and (ii) ensure quality-control (e.g., light-touch editorial review whereby posted information appears as "unverified" until an editor reviews and approves it).

Ironing out these conceptual, epistemological, and statistical issues is a pre-requisite for setting out to build an actual web interface that researchers can use to track and gauge replicability and generalizability to ultimately produce a living and dynamically evolving body of **knowledge** that can soundly inform public-policy. The framework is developed with an initial focus on social science findings that have applied implications given that such findings have a lot more potential in influencing society (e.g., [font disfluency boosts math performance effect](#); stereotype threat; ["wise" interventions on voting](#); Mozart effect). That said, the framework will also be able to handle some basic social science findings that may not necessarily have direct societal implications.

### 1. Accommodating Different Approaches to Replications

There are currently different approaches to – and terminologies associated with – replication both within and across social sciences fields (Goodman, Fanelli, & Ioannidis, 2016). Replication in economics and political science typically entails attempts to independently reproduce originally reported results by repeating the same analyses on the raw (or transformed) data, a practice dubbed “pure replication” (Hamermesh, 2007; what we’re calling “analytic reproducibility”<sup>2</sup>). A “replication” in this context often also includes “robustness analyses” that attempt to reproduce originally reported results using different – and often more justifiable/valid – statistical analyses and models (Christensen, 2016; sometimes also called “sensitivity checks”). If originally reported results can be successfully reproduced (from the data) using the same statistical analyses, and if results are robust to different analyses and models, then our confidence in the reported results is bolstered. Consequently, this potentially justifies following-up the research to see if the finding replicates in a new independent sample using similar methodology. In contrast, if results cannot be reproduced using the same analyses, or if results are only reproducible using similar analyses but are not robust via other more valid statistical approaches and models, then our confidence in the original results is reduced/tempered. Consequently, following-up on the findings via more expensive independent sample replications may not be worth the required resources, depending on the nature of the research.

---

<sup>2</sup> We use the “analytic reproducibility” terminology because (1) it disambiguates the activity from the traditional notion of (new sample) replication in science and (2) it more directly and concretely describes the actual activity.

For example, in probably the most well-known “replication” in economics, Herndon, Ash, and Pollin (2014) attempted to replicate Reinhart and Rogoff’s (2010) main finding that public debt/GDP ratios above 90% consistently reduce a country’s GDP growth, using the same dataset. Herndon et al. were unable to analytically reproduce Reinhart and Rogoff’s main finding (using similar statistical approaches) and also found that the original finding was not robust across alternative reasonable methods of calculation, due to coding errors, inappropriate weighting methods, and selective exclusion of available data. These replication results refuted Reinhart and Rogoff’s main finding and had major implications in questioning support of austerity policies in the aftermath of the 2007-2009 financial crisis, which Reinhart and Rogoff’s findings influenced.

In (experimental) psychology, however, replication has typically been approached very differently, with a near-exclusive focus on attempting to replicate original findings in new independent samples using methodology similar to an original study. This is likely due to the inexpensive nature of small-scale lab experiments and online survey studies, compared to the more expensive representative and/or as large as possible samples typically used in economics and political science (Christensen, 2016). For example, the influential social psychology finding dubbed the “Macbeth effect” (Zhong & Liljenquist, 2006, *Science*), whereby a threat to one’s moral purity induces the need to cleanse oneself, has been the target of several direct replication attempts by several different labs collecting new samples of data (e.g., Earp, Everett, Madva, & Hamlin, 2014; Fayard, Bassi, Bernstein, & Roberts, 2009; Gamez, Diaz, & Marrero, 2011). Across three distinct operationalizations of the “Macbeth effect” hypothesis (Zhong & Liljenquist’s Study 2, Study 3, and Study 4), independent labs have been unable to replicate the original findings, with meta-analytic estimates of replication effect sizes for each operationalization not statistically significantly different from zero (see Figure 2 below). As this example demonstrates, the primary difference in the approach to replication adopted by psychology (compared to economics and political science) is not just a focus on re-testing an original hypothesis using similar methodology via the collection of new data, but also by executing new sample direct replications across *distinct operationalizations* of an original hypothesis.<sup>3</sup>

---

<sup>3</sup> Exceptions include Innovations for Poverty Action’s (<http://www.poverty-action.org/>) impact evaluations and AidGrade.org’s meta-analyses of international development efforts, however, these initiatives do not meta-analyze evidence from replication studies (which typically don’t yet exist) but do so using traditional meta-analytic approaches which we now know cannot properly account for publication bias and analytic and design flexibility problems that afflict each original study included in a traditional meta-analysis (see Replication Taxonomy section for a detailed elaboration on how curating replication studies employing sufficiently similar methodologies eliminate/minimize such biases).

By accommodating and unifying the different approaches to replication within the social sciences, the proposed framework helps us achieve our goal of accumulating trustworthy knowledge that can soundly inform public policy by identifying replicability and generalizability gaps that currently exist in the different social science fields. That is, psychology has done well in terms of replicability in independent samples, but needs to maximize efficiency of research resources by first probing and ensuring analytical reproducibility and robustness before attempting resource- and time-costly independent sample replications. On the other hand, economics and political science have done well in efficiently probing reproducibility and robustness, but need increased focus, when possible, on replicability and generalizability via independent samples for findings that have “survived” the reproducibility and robustness verification hurdles. Sometimes independent samples *are* used, but these are typically *pre-existing* datasets that typically speak more to generalizability than replicability given the datasets were generated by studies using different methodologies, populations, and/or domains.

Indeed, such gaps in gauging the replicability and generalizability of social science findings is reflected in the different web platforms currently used to track and organize replications across the three main social science disciplines:

- **Economics:** [ReplicationWiki](#) is a wiki-based platform linking published “replication articles” (and unpublished PhD student seminar replication papers) to original articles.
- **Political Science:** [Harvard Dataverse repository](#) of unpublished “replication articles” of original articles.
- **Psychology:** [PsychFileDrawer.org](#) a repository of unpublished reports of (mostly) direct replications of original findings (some of these replications have now been published as part of peer-reviewed journal articles).

Each of these platforms suffers from various deficiencies and suboptimalities. For example, the [PsychFileDrawer.org](#) platform doesn’t allow the posting of analytic reproducibility and robustness checks and also only includes unpublished (but not published) reports of new sample independent replications. The replication repositories in economics and political science both don’t allow linking distinct (new data) replications of an original study to each other, nor do they allow meta-analytic synthesis of such replications to quantitatively gauge replicability and generalizability ([PsychFileDrawer.org](#) also doesn’t allow linking and meta-analyzing replication results). The proposed unified framework overcomes all of these shortcomings and simultaneously yields the additional benefit of inter-disciplinary curation of social science evidence, which maximizes the creation of trustworthy and actionable knowledge relative to the total costs of producing all of the currently available empirical evidence.

## 2. Accommodate Different Kinds of Evidence Collections

It should be noted that the different approaches to replication just reviewed involve different units of analysis. Reproducibility and robustness verifications are done at the study-level nested within articles whereby “replication” articles are directly connected to original articles. On the other hand, while independent sample replications and generalizability studies are also at the study-level, the focus is on curating replication evidence for specific effects/hypotheses across different generalizations and/or operationalizations of the target constructs, speaking to the replicability and generalizability of an effect. Hence, we need different kinds of what we’re calling “evidence collections” to account for the fact that replication studies of specific effects/hypotheses can be nested in different ways in relation to original studies that tested an effect across different generalizations and/or operationalizations of the target constructs. Evidence collections need to accommodate:

- *replications of an effect via a single generalization/operationalization originating from a single published article*
- *replications of an effect via multiple generalizations/operationalizations originating from a single published article*
- *replications of an effect via multiple generalizations/operationalizations originating from multiple published articles*

Yet another kind of evidence collection includes replications of *sets of effects/hypotheses* (across different generalizations and/or operationalizations) derived from a specific theory that originate from multiple published articles. Such evidence collections are beyond the scope of the current framework and hence will not be further discussed.

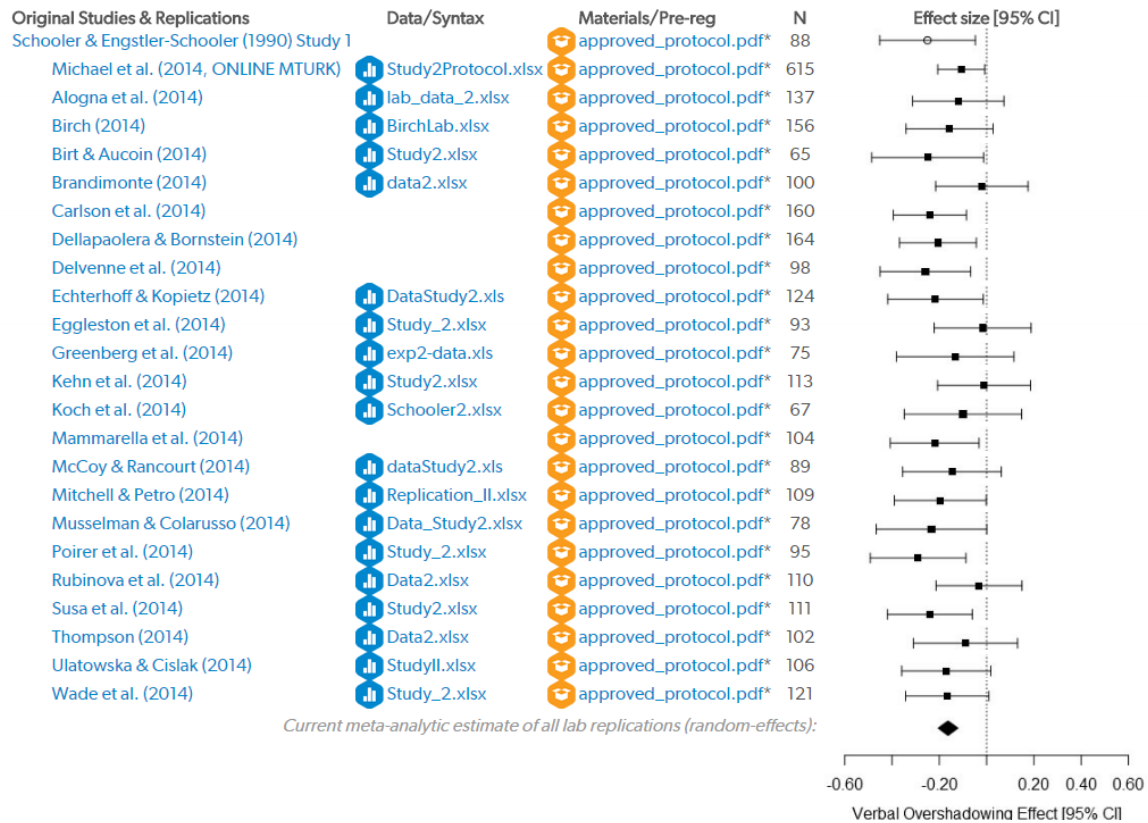
An example of the first kind of evidence collection (*replications of an effect via a single generalization/operationalization originating from a single published article*) can be seen in Figure 1 which shows the many replications of Schooler and Engstler-Schooler’s (1990, Study 1) verbal overshadowing effect, which were part of the first multi-lab Registered Replication Report (RRR) initiative spearheaded by *Perspectives on Psychological Science* (Simons & Holcombe, 2013).

## Schooler & Engstler-Schooler (1990) -- Replications (23) [Tweet](#) [Link](#)

Verbal overshadowing of visual memories: Some things are better left unsaid

DOI:10.1016/0010-0285(90)90003-M 

[Original Abstract]



**Figure 1:** Replications of Schooler and Engstler-Schooler's (1990, Study 1) verbal overshadowing effect as an example of the first kind of evidence collection whereby replications are curated for an effect via a single generalization/operationalization originating from a single published article.

An example of the second type of evidence collection (*replications of an effect via multiple generalizations/operationalizations originating from a single published article*) can be seen in Figure 2 which shows different direct replications of Zhong and Liljenquist's (2006) original Study 2, 3, and 4 of the "Macbeth effect", whereby a threat to one's moral purity (IV) was hypothesized to induce the need to cleanse oneself (DV) as a way to "wash away one's sins". The first generalization of the "Macbeth effect" (Zhong & Liljenquist's Study 2) involved testing whether transcribing a first-person description of an unethical vs. ethical act (IV) increased the need to cleanse oneself via the desirability of cleansing products (DV). The second generalization of the "Macbeth effect" (Zhong & Liljenquist's Study 3) involved testing whether recalling an unethical vs. ethical deed from one's own past (IV) increased likelihood of choosing an antiseptic wipe (rather than a pencil) as a free gift (DV). The third generalization of the "Macbeth effect" (Zhong & Liljenquist's Study 4) involved having all participants recall an unethical deed and testing

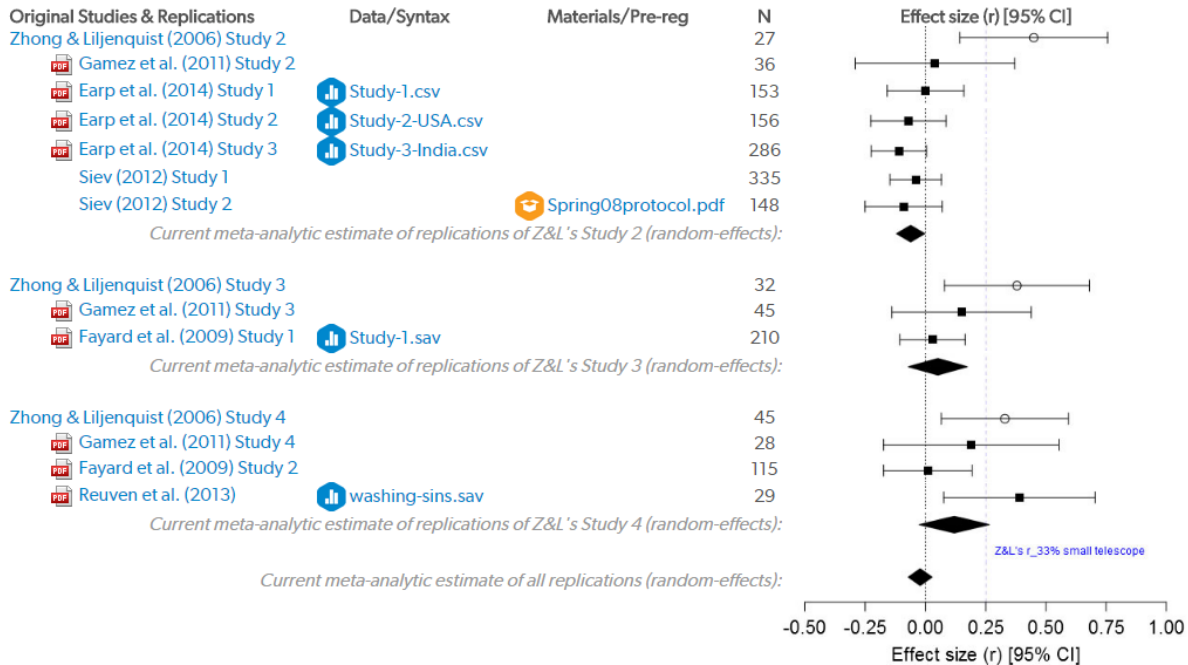
whether cleansing one's hands with an antiseptic wipe vs. not (IV) decreased the likelihood of volunteering to participate in another study without pay (an altruistic gesture) (DV). As can be seen in Figure 2, replication difficulties have subsequently emerged for each of the three generalizations of the “Macbeth effect”.

## Zhong & Liljenquist (2006) -- Replications (11) [Tweet](#) [Link](#)

Washing away your sins: Threatened morality and physical cleansing

DOI:10.1126/science.1130726 [PDF](#)

[Original Abstract]



**Figure 2:** Replications of Zhong and Liljenquist's (2006) original Study 2, 3, and 4 of the “Macbeth effect” as an example of the second kind of evidence collection whereby replications are curated for an effect via multiple generalizations/operationalizations originating from a single published article.

An example of the third type of evidence collection (*replications of an effect via multiple generalizations/operationalizations originating from multiple published articles*) can be seen in Figure 3 which shows different replications of the “money priming” effect originally reported across distinct original studies in two different published articles (Vohs, Mead, & Goode, 2006; Caruso, Vohs, Baxter, & Waytz, 2013). Direct replication evidence has emerged for five generalizations of the “money priming” effect, one generalization from Vohs et al.'s Study 3 and four generalizations from Caruso et al.'s Study 1, 2, 3, and 4.



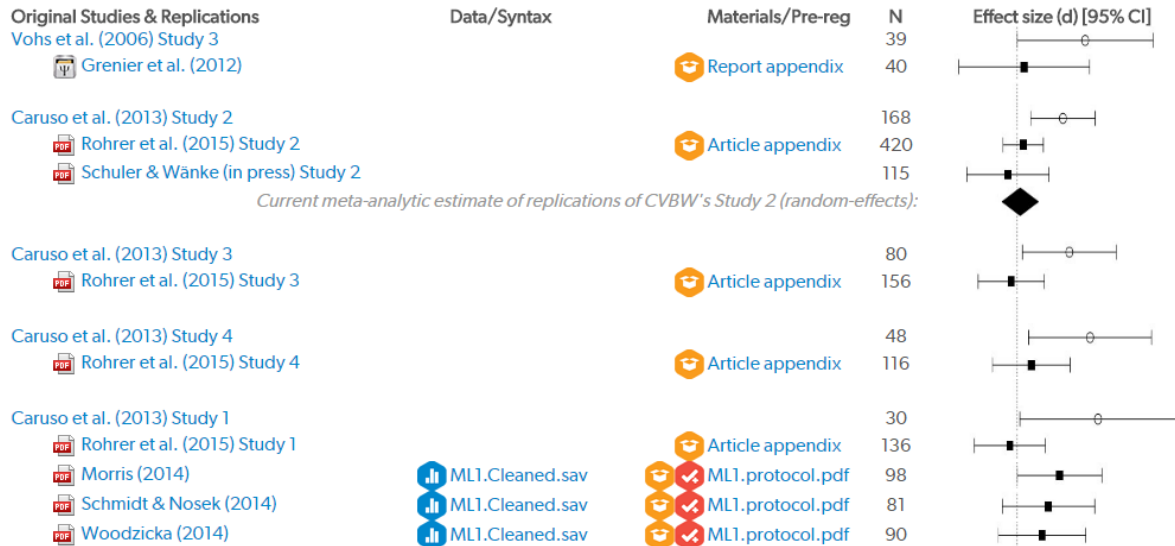
## Money priming -- Replications (42) [Tweet](#) [Link](#)

Vohs, Mead, & Goode (2006) [PDF](#)

The psychological consequences of money

Caruso, Vohs, Baxter, & Waytz (2013) [PDF](#)

Mere exposure to money increases endorsement of free-market systems and social inequality



**Figure 3:** Replications of Vohs, Mead, and Goode's (2006) Study 3 and Caruso, Vohs, Baxter, and Waytz' (2013) Study 1, 2, 3, and 4 as an example of the third kind of evidence collection whereby replications are curated for an effect via multiple generalizations/operationalizations originating from multiple published articles.

As can be seen, the different types of evidence collections involve different kinds of nested parent-child (original-replication study) pairings to account for the fact that an effect/hypothesis is typically tested via different operationalizations of the constructs and/or generalizations to rule out method artifacts and validity issues (Campbell & Fiske, 1959). In addition to gauging replicability in a nuanced manner, such an accommodating framework also speaks to the generalizability of an effect given that it documents replication evidence showing an effect generalizes to other methodologies, contexts/domains, and populations, which are crucial for achieving our goal of creating trustworthy knowledge that can soundly inform public-policy.

A final note to mention regarding the flexibility of the proposed framework is that a specific original or replication study could be simultaneously included in different evidence collections depending on the nature of the effect/hypothesis for which replication evidence is being curated. For example, replications of Bargh and Shalev's (2012) Study 1 and 2 Loneliness-hot-shower effect could simultaneously be included in a distinct but broader evidence collection curating replication evidence of embodiment of warmth effects more generally, in addition to

being included in a more narrow evidence collection focusing specifically on the loneliness-hot-shower effect.

### 3. Accommodate Different Kinds of Study Designs

The proposed conceptual framework also needs to accommodate different kinds of study designs for studies included in evidence collections. At a broad level, we need to distinguish between experimental study designs (e.g., randomized controlled trials [RCTs]) and observational study designs (typically called “correlational designs” in psychology). Among experimental study designs, at minimum we need to accommodate between-subject designs, within-subject designs, and mixed between-and-within designs (and nested within each of these are simple two- or three-condition designs or more complex interaction designs). Among observational designs, we need to accommodate simple correlational designs, more complex regression-based designs, and even more complex structural equation (SEM) models.

It’s important to note that our initial efforts will be focused on the curation of replication evidence of original findings using *experimental designs*, given these are considered the gold standard for producing *causal evidence*, which consequently have much more potential in informing public-policy decisions. Naturally, initial curation efforts will first be focused on the simplest two- or three-condition between-subject and within-subject designs, which are among the most commonly employed in the social sciences. Subsequent efforts will then focus on more complex interaction and mixed designs and observational/correlational designs.

### 4. Replication Taxonomy

Our proposed conceptual framework also needs some kind of workable replication taxonomy to allow a principled and justifiable approach to distinguishing replications that are sufficiently methodologically similar to an original study from replications that are insufficiently similar. Contrary to some current views in the field of psychology (e.g., Finkel, Eastwick, & Reis, 2015; Crandall & Sherman, 2016), replications actually lie on an ordered continuum of methodological similarity relative to an original study, with exact and conceptual replications occupying the extremes. A direct replication repeats a study using methods as similar as is reasonably possible to the original study, whereas a conceptual replication repeats a study using different general methodology and tests whether a finding generalizes to different manipulations, measurements, domains, and/or contexts (Asendorpf et al., 2013; Brandt et al., 2014; Lykken, 1968; Simons, 2014). To guide the classification of replications based on methodological similarity to an original study, we use the replication taxonomy depicted in Figure 4, which is a simplification of Schmidt’s (2009) replication classification scheme, itself a

simplification of an earlier taxonomy proposed by Hendrick (1991) (see also Hüffmeier, Mazei, & Schultze, 2016).

Design facet	Direct replication			Conceptual replication	
	Exact replication (Everything controllable the same)	Very close replication (Procedure or context is different)	Close replication (IV or DV stimuli are different)	Far replication (IV or DV operationalization is different)	Very Far replication (Everything can be different)
IV operationalization	same	same	same	different	
DV operationalization	same	same	same	different	
IV stimuli	same	same	different		
DV stimuli	same	same	different		
Procedural details	same	different			
Physical setting	same	different			
Contextual variables	different				
⋮	⋮				

**Figure 4:** Replication taxonomy to guide the classification of relative methodological similarity of a replication study to an original study. “Same” (“different”) indicates the design facet in question is the same (different) compared to an original study. Proposed framework treats “Exact”, “Very Close”, and “Close” replications as sufficiently methodologically similar to an original study to count as a direct replication. “Far” replications are only considered in the sense that direct replication evidence can be curated across different operationalizations of an effect.

As can be seen in Figure 4, different types of increasingly methodologically dissimilar replications exist between these two poles, each of which serve different purposes. In an “Exact” replication (1<sup>st</sup> column), every controllable methodological facet would be the same except for contextual variables, which is only typically possible for the original lab and hence is of limited utility for our purposes here. “Very Close” replications (2<sup>nd</sup> column) employ the same IV and DV operationalizations *and* IV and DV stimuli as an original study, but can be different in terms of procedural details, physical setting, and contextual variables.<sup>4</sup> “Close” replications (3<sup>rd</sup> column) employ the same IV and DV operationalizations, but can employ different sets of IV or DV stimuli (or different scale items or a shorter version of a scale) and different procedural details and contextual factors. “Far” replications (4<sup>th</sup> column) involve different operationalizations for the IV or DV constructs whereas for “Very Far” replications (5<sup>th</sup> column) *everything can be different* including different constructs altogether (via different operationalizations) in different domains of inquiry (as in Bargh, Chen, & Burrows’, 1996 Study 1, 2, and 3). Hence, “Exact”, “Very Close”, and “Close” replications reflect increasingly methodologically dissimilar types of direct replications that provide sufficient levels of falsifiability to (1) test and confirm the reliability (i.e., basic existence) of a phenomenon and (2) systematically test relevant contextual factors

<sup>4</sup> With any required linguistic and/or cultural adaptations of the IV or DV stimuli considered as part of “Contextual variables.”

and other auxiliary assumptions, which contribute to validity and generalizability (Srivastava, 2014; Meehl, 1967, 1978; see also LeBel, Berger, Campbell, & Loving, 2016). On the other hand, “Far” and “Very Far” replications can only speak to validity and generalizability, given the major design differences intentionally introduced (and only pursuing such follow-up studies is dangerous, more on this below).

To achieve our goal, only the three types of direct replications (“Exact”, “Very Close”, and “Close”) are eligible for inclusion in evidence collections. We need such a demarcation because sufficiently methodologically similar replications naturally constrain design and analytic flexibility (old-school “poor person’s pre-registration”) and so ensures sufficient levels of falsifiability to refute an original claim, assuming auxiliary assumptions are met (Earp & Trafimow, 2015; Meehl, 1967, 1978). If a follow-up study is different on every methodological facet, then it can never refute an original claim because unsupportive evidence can always be attributed to one of the intentionally introduced design differences rather than to the falsity of the original hypothesis (Hendrick, 1991; LeBel & Peters, 2011; see also Feynman, 1974). Without such constraints, a popular field where numerous researchers are testing the same (false) hypothesis will inevitably produce false positives with enough determination given that typically an infinite number of analytic and design specifications (across different operationalizations) exist to test a specific effect/phenomenon (Ioannidis, 2005; 2012). The historical case of cold fusion provides a compelling example of this. As recounted by Pashler and Harris (2012, p. 534), only follow-up studies using very different methodology yielded a trickle of positive results observing a cold fusion effect whereas more methodologically similar replications yielded overwhelmingly negative results (Taubes & Bond, 1993).

Ambiguous edge cases may sometimes occur where it is difficult to determine whether a “Close” replication is sufficiently methodologically similar to an original study. In these cases, we will err on the side of being overly exclusive to minimize bias, however, ambiguous replications will nonetheless be listed in evidence collections, but unchecked/excluded by default from the meta-analytic estimate. Crucially, design differences will be prominently displayed to allow readers to make up their own mind as to whether such design deviations should even matter (Simonsohn, 2016). Users could then see how including or excluding specific replications alters the meta-analytic result (similar to how we deal with eligible direct replications from non-independent researchers, see next section).

It’s important to note that “Far” replications are only included in evidence collections as separate original study generalization branches under which “Exact”, “Very Close”, and “Close” replication evidence is curated (e.g., Zhong & Liljenquist’s, 2006, Study 3 can be considered a

“Far” replication of Zhong & Liljenquist’s, 2006, Study 2, given it tests the same hypothesis via different operationalizations of the target constructs). The framework could be extended to also allow the curation of direct replication evidence of “Conceptual” replications as separate original study generalization branches, however, as a first step, we will only consider the simpler case where generalization branches involve “Far” replications.

### 5. Accounting for Replication Study Quality

Not all replication studies are created equally. Consequently, our proposed conceptual framework also needs to take into account the *quality of replication studies*, so that readers can easily integrate this information to facilitate more accurate interpretation of replication results. At minimum, however, replication studies (like any study) need to have sufficient reporting of methodological details. The following study-level information is required to be eligible for inclusion in an evidence collection: IVs and their operationalizations, DVs and their operationalizations, design (e.g., between-subjects), methods protocol (whether included in a published article or unpublished report [e.g., as is the case on [PsychFileDrawer.org](https://PsychFileDrawer.org)]), sample size (N), and effect size, with the first three required fields needed to determine whether a replication is sufficiently methodologically similar (as described in the previous section). It will be assumed that all replication studies have abided by the Basic 4 reporting standards (i.e., disclosure of all excluded observations, tested experimental conditions, assessed outcome measures, and sample size determination rule), inspired by Simmons et al.’s (2012) 21-word disclosure solution and popularized by PsychDisclosure.org (LeBel et al., 2013), and now adopted by Psychological Science (Eich, 2014) and dozens of other journals in the social sciences (LeBel & John, in press). In the future, we may consider more comprehensive and formal reporting standards that replication studies would need to abide by (e.g., CONSORT reporting standards for RCTs, Schulz, Altman, & Moher, 2010; STROBE reporting standards for observational studies, Vandenbroucke et al., 2014).<sup>5</sup>

In terms of gauging replication study quality, six (optional) study features can be considered and should be considered if they are available: (1) verifiability, (2) pre-registration status, (3) analytic reproducibility verification status, (4) analytic robustness verification status, (5) active sample evidence, and (6) design differences. First, *verifiability* can be used as an indicator of study quality in the sense that – all else being equal – studies with links to open data and open materials are likely of higher quality, even if the data and materials have not been formally verified (knowing one’s data and materials are fully independently verifiable induces

---

<sup>5</sup> It should be noted that evidence collections will themselves be reported in a comprehensively transparent manner following the PRISMA guidelines for reporting systematic reviews and meta-analyses (Moher, Liberati, Tetzlaff, & Altman, 2009).


more careful quality controls during design and analysis). Second, replication studies that have been pre-registered (e.g., on the [OSF](#) or at [AsPredicted.org](#)) can be considered higher quality in the sense that analytic and design flexibility are even more constrained than non-pre-registered replication studies, wherein minor analytic and design flexibility may still exist. Third, the analytical reproducibility verification status of a replication study is indicative of study quality given that, all else being equal, one can justifiably have more confidence in a study if an independent researcher has endorsed a replication study's results as analytically reproducible from publicly available open data (as is the case in endorsing the analytical reproducibility of an *original study*). Fourth, similar reasoning applies to studies that have been endorsed as analytically robust to alternative (and often more justifiable and/or valid) statistical approaches and models.





Fifth, replication studies that provide information regarding *active sample evidence* (or what others have called *positive controls*, e.g., Cusack, Vezenkova, Gottschalk & Calin-Jageman, 2015; Moery & Calin-Jageman, 2016) can be considered higher quality given that one can be more confident that the replication study's instruments were "active" in being able to detect the sought after effect, independent of the results of the study's main outcome. This is particularly important when interpreting replications reporting null results because active sample evidence rules out that null findings are due to an anomalous sample and/or fatal data processing errors. For example, successful manipulation checks or replicating a past known gender difference on the outcome variable each help establish that the replication study had active ability to detect the original finding beyond being sufficiently powered.<sup>6</sup> Figure 5 demonstrates an example of active sample evidence whereby mean ratings of nudes were rated as more pleasant than mean ratings of abstract art photos in Balzarini et al.'s (2015) unsuccessful direct replication attempts of Kenrick et al.'s (1989) playboy effect.

---

<sup>6</sup> We've now moved away from requiring replication studies to have at least 70-80% a priori power (as in Curate Science Version 1.0.5) given our current focus on meta-analytic thinking.

## Playboy Effect: Kenrick, Gutierrez, & Goldberg (1989) -- Replications (3) [Tweet](#) [Link](#)

Influence of Popular Erotica on Judgments of Strangers and Mates 

Results/Open Data	Design Details			
Original Studies & Replications	Independent Variables	Dependent Variables	Design Differences	Active Sample Evidence
 Kenrick et al. (1989) Study 2	Playboy centerfolds vs. control Participant sex	Love for partner (Rubin Love-scale)	-	
 Balzarini et al. (2015) Study 1	Playboy centerfolds vs. control Participant sex	Love for partner (Rubin Love-scale)	Updated pictures of abstract art & male/female nudes (as suggested by Kenrick) Two attention check questions (online sample)	Nudes rated as more pleasant than abstract art
 Balzarini et al. (2015) Study 2	Playboy centerfolds vs. control Participant sex	Love for partner (Rubin Love-scale)	Updated pictures of abstract art & male/female nudes (as suggested by Kenrick) Two attention check questions (online sample)	Nudes rated as more pleasant than abstract art
 Balzarini et al. (2015) Study 3	Playboy centerfolds vs. control Participant sex	Love for partner (Rubin Love-scale)	Updated pictures of abstract art & male/female nudes (as suggested by Kenrick) Two attention check questions (online sample)	Nudes rated as more pleasant than abstract art

**Figure 5:** Mean ratings of nudes were rated as more pleasant than mean ratings of abstract art photos (right-most column) in Balzarini et al.'s (2015) unsuccessful direct replication attempts of Kenrick et al.'s (1989) playboy effect, providing positive control evidence.

A final aspect of replication studies to consider (though not strictly about study quality, but nonetheless related) is replication study design differences that deviate from the original study, which have recently been argued deserve increased attention (Brandt et al., 2014). As stated by [Simonsohn \(2016\)](#), prominently disclosing design differences “better enables readers to consider the consequences of such differences, while encouraging replicators to anticipate and address, before publication, any concerns they may raise.”

With all such information prominently and easily accessible, a reader will be better able to take into account replication study quality when assessing overall replicability across replications by intuitively weighting the replications according to the aforementioned study quality dimensions, but also by quantitatively pooling across different subsets of replications that vary on the study quality dimensions by toggling them on-or-off for inclusion in the meta-analytic calculations.

The framework also accounts for replication study quality in relation to the problem of “correlated investigators” (Rosenthal, 1991). Consistent with the independent corroboration feature of the scientific spirit, only eligible direct replications that are *independent* (meaning no overlap in authors) are included in the formal meta-analytic estimates of replication effect sizes. That said, eligible non-independent direct replications can nonetheless be added in an evidence collection and toggled on-or-off for inclusion in meta-analytic calculations for exploratory purposes.

## 6. Meta-analytic Synthesis of Replication Evidence

Our general framework also requires a principled statistical approach to (1) meta-analytically combine replication evidence to yield mean effect size estimates for each effect generalization and an overall mean effect size estimate across generalizations and (2) to interpret these meta-analytic results, and to do so in a sequential cumulative manner as in *continuously cumulating meta-analytic* approaches (CCMA, Braver, Thoemmes, & Rosenthal, 2014; see also Borenstein, Hedges, Higgins, & Rothstein, 2009). As is standard in traditional meta-analyses, combined effect size estimates should be weighted by the precision of each individual replication study, with studies reporting more precise estimates given proportionally more weight. The precise nature of this weighting, however, depends crucially on the nature of the studies being combined, which dictates whether a fixed-effect model or random effects model should be implemented (Borenstein, Hedges, & Rothstein, 2007).

A fixed-effect model (or common-effect model) assumes the same true effect size is being tested across (replication) studies and the combined effect is simply the mean of the study effect sizes weighted by the precision of each study. On the other hand, a random-effects model allows for the possibility that the true effect size may vary across (replication) studies (e.g., due to different methodologies or different subject characteristics from different populations) and the combined effect is the mean of the *distribution* of study effect sizes, also weighted by each study's precision.

Given that direct replications (of an original study) eligible for curation within our framework are allowed to differ methodologically in minor ways from other direct replications (i.e., "Close" vs. "Very Close" vs. "Exact", see taxonomy above), it follows that in general a random-effects model will be more appropriate than a fixed-effect model (which will be implemented via the popular `metafor` R package; Viechtbauer, 2015). Furthermore, even in the case where all direct replications are in the same replication category (e.g., "Very Close"), we may still not necessarily expect the same true effect size to arise across replications given linguistic and/or cultural differences across studies, depending on the nature of the effect (e.g., social psychological phenomena could be expected to be more linguistically and/or culturally-variable than cognitive or bio-physiological phenomena; see for e.g., Trafimow & Earp, 2016). In these cases, a random-effects model may still be the most appropriate model to implement.

In cases where replications are meta-analyzed across multiple generalizations, a hierarchical (or multi-level) random-effects approach should be taken whereby replications are nested within generalizations, which are treated as a random factor. This is to allow for the possibility that a true effect size may vary within *and* across generalizations and also to account for the fact that replication effect sizes within a generalization are expected to be more similar to



each other than across generalizations. In the more complicated cases where replications include both a primary and secondary outcome, an additional level would be added to account for the correlated nature of the outcome variables.

A random-effects model approach can also estimate heterogeneity across replication effect sizes by testing whether variability in effect size estimates across replications is greater than can be attributable to sampling error alone. This can be useful for suggesting the possibility that replication effect sizes may systematically vary depending on unknown or unmeasured variables (e.g., subject-characteristics or contextual or “expertise” factors). That said, not finding evidence for heterogeneity should not be interpreted as evidence for homogeneous effect sizes across replications given that perhaps a small amount of heterogeneity exists but there was insufficient power to detect it (hence, a larger number of studies would be required to reliably detect such small amount of heterogeneity). Hence, when the number of replication studies is very small, it may not be possible to estimate between-study heterogeneity and so one should remain agnostic regarding the heterogeneity/homogeneity of effect sizes.

Concerning the *interpretation of overall meta-analytic results*, the general approach will be to examine confidence intervals (CIs) of meta-analytic effect size estimates within each effect generalization (when multiple replications are available within a generalization) and CIs of the overall meta-analytic effect size estimate. Meta-analytic effect sizes with CIs excluding 0 in the same direction as an original study can be interpreted as suggesting a reliable effect, which should be examined for each generalization separately. Such frequentist confidence intervals, however, are limited in that they don't indicate the range of likely values of the true effect size but rather indicate that in the hypothetical long-term, 95% of CIs calculated will include the population true effect size. To overcome this limitation, we plan to implement Bayesian meta-analytic approaches that yield principled interval ranges via credible intervals (e.g., [Ding & Baio, 2016](#); Edwards, Lindman, & Savage, 1963; Guan, & Vandekerckhove, 2016) or highest-density intervals (Kruschke, 2013).

Independent of the type of confidence interval used, it's important to avoid interpreting the absence of evidence as evidence that a phenomenon/effect doesn't exist given that it's possible the phenomenon is in fact replicable using different operationalizations and/or improved measurement instruments. This is the case even if a phenomenon appears non-replicable across several different generalizations. That said, given a certain amount of unfavorable replication evidence across a certain number of generalizations, the community of researchers may come to consensually agree that continued research on testing a specific phenomenon may be an unwise use of funding. This will depend on the predicted technological

developments of the measurement instruments required to empirically assess the target phenomenon, but also the specific utility and disutility of the different research outcomes involved (Miller & Ulrich, 2016; LeBel et al., 2016).

In cases where only a few replication studies are available (e.g.,  $k = 2$  to 4 studies), Simonsohn's (2015) small telescope approach may be useful in augmenting the interpretation of the limited meta-analytic estimates (either within or across generalizations). That is, assessing the extent to which replication results are consistent with an effect size big enough to have been detectable in the original study (effect size original study had power = 33% to detect). Another approach to use in this situation is to display the minimum effect size reliably detectable (@ 95% power) based on the overall total sample size of all replications (e.g.,  $d_{\min_{95\text{power}}} = .20$ ). One could then conclude that a true effect size may still exist that is smaller than this minimum effect size, but that significantly more resources would have to be invested. Implications of such results could then be determined in relation to the utility and disutility of relevant research outcomes in question (Miller & Ulrich, 2016).

Finally, to maximize the interpretability of replication results, generalization-specific unstandardized effect size estimates (tied as directly as possible to the outcome variable assessed) will be used within each generalization (as done in meta-analyzing multi-lab replications in Registered Replication Reports (RRRs) at *Perspectives on Psychological Science*; Simons & Holcombe, 2013), while overall meta-analytic effect size estimates will be scored in a consistent direction and presented in standardized units (e.g., z-score).

## 7. Viable Crowd-sourcing System

As a final step, a viable crowd-sourcing system is needed to allow the community of researchers to curate replication evidence following our general conceptual framework via an easy-to-use web platform. This platform is crucial because scientific evidence is dynamic and constantly evolving: New evidence can always refute an accepted hypothesis and likewise new ways of testing a previously discarded hypothesis (e.g., via improved measurements or designs) can lead to corroboration of that hypothesis. In this way, it simply no longer makes sense to continue publishing literature reviews of evidence (as in traditional meta-analyses) within a static document that can literally become out-of-date a few days after such document is submitted for peer-review to a scientific journal. Key features of such web platform will include:

- Features to incentivize number and frequency of contributions:
  - Each user's various contributions (e.g., adding replications, curating replication information, commenting, etc.) will be prominently summarized and highlighted on their public profile page as a way to reward contributions.

- The home page will also feature recently added and updated evidence collections and prominently display the user(s) who contributed the new information.
- Follow a “low barrier to entry” (incremental) approach to maximize the number and frequency of contributions. This will be achieved by having as many “optional” fields as possible, so that editors and users can subsequently continue and finish curation at a later point in time.
- Light-touch editorial review to ensure sufficient quality control. For example, when a registered researcher posts a new replication study to an existing evidence collection, the information will appear as “unverified” until an editor reviews and approves it. During the initial phase of curation, only editors will be allowed to create new evidence collections. That said, our vision is that eventually registered users will also be able to create new evidence collections, which will appear as “unverified” until an editor approves it.

As just described, the general framework is very flexible in overcoming several distinct conceptual, epistemological, and statistical challenges that arise to track and gauge replicability and generalizability of empirical social science findings over time. This unified replication evidence curation framework can now guide the design of a user interface (UI) required to build an actual web platform that researchers can use to track and interpret the replicability and generalizability of social science findings. Please see Appendix A (currently in development; near completion) for prototype UI wireframes of all major pages and views (i.e., home page, search results page, evidence collection page [view & edit modes], article page, study page [including data viewer, analytic reproducibility and analytic robustness windows], researcher page, user profile page, and administrator dashboard page).

### **Conclusion**

Every year, society invests billions of dollars funding scientific studies to deepen our understanding of the natural and social world, with the hope the findings yielded by these studies help us overcome and/or address important societal problems. Scientific findings, however, are not reliable knowledge. Rather, reliable knowledge emerges via the systematic accrual -- and nuanced interpretation -- of replicability and generalizability evidence. For this, we need to track and gauge the replicability and generalizability of empirical effects over time in a systematic and gradual fashion. The proposed general replication evidence curation framework guides the design of a web platform to allow social science researchers to do just this.

Systematically tracking and interpreting the replicability and generalizability of empirical effects over time will allow the social science community to create a dynamically living and continuously evolving body of reliable **knowledge** that can soundly inform public-policy to improve and solve important societal problems.

## References

- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108-119.
- Balzarini, R. N., Dobson, K., Chin, K. A., & Campbell, L. (2016, August 18). Independent Replication of Kenrick, Gutierrez, & Goldberg (1989, JESP, Study 2). Retrieved from <http://osf.io/njzmq>
- Bargh, J. A., & Shalev, I. (2012). The substitutability of physical and social warmth in daily life. *Emotion, 12*(1), 154-162.
- 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*(2), 230.
- Borenstein, M., Hedges, L., & Rothstein, D. (2007). Meta-analysis: Fixed effect vs. random effects. Retrieved from [http://www.metaanalysis.com/downloads/Meta-analysis\\_fixed\\_effect\\_vs\\_random\\_effects\\_sv.pdf](http://www.metaanalysis.com/downloads/Meta-analysis_fixed_effect_vs_random_effects_sv.pdf)
- Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). Cumulative Meta-Analysis. In M. Borenstein, L. V. Hedges, J. Higgins, & H. R. Rothstein (Eds.), *Introduction to Meta-Analysis* (371-376). John Wiley & Sons.
- Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication?. *Journal of Experimental Social Psychology, 50*, 217-224.
- Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*(3), 333-342.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin, 56*(2), 81.
- Caruso, E. M., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General, 142*(2), 301-306.
- Christensen, G. (2016). Manual of best practices in transparent social science research. Retrieved from <https://github.com/garretchristensen/BestPracticesManual/blob/master/Manual.pdf>
- Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology.*
- Cusack, M., Vezenkova, N., Gottschalk, C., & Calin-Jageman, R. J. (2015). Direct and Conceptual Replications of Burgmer & Englich (2012): Power may have little to no effect on motor performance. *PloS one, 10*(11), e0140806.
- Ding, T., & Baio, G. (2016). Package 'bmeta'. The Comprehensive R Archive Network. Retrieved from <https://cran.r-project.org/web/packages/bmeta/bmeta.pdf>

- Earp, B. D., Everett, J. A., Madva, E. N., & Hamlin, J. K. (2014). Out, damned spot: can the “Macbeth Effect” be replicated?. *Basic and Applied Social Psychology*, 36(1), 91-98.
- Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology*, 6, 621.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference in psychological research. *Psychological Review*, 70, 193-242
- Eich, E. (2014). Business not as usual. *Psychological Science*, 25(1), 3-6.
- Fayard, J. V., Bassi, A. K., Bernstein, D. M., & Roberts, B. W. (2009). Is cleanliness next to godliness? Dispelling old wives’ tales: Failure to replicate Zhong and Liljenquist (2006). *Journal of Articles in Support of the Null Hypothesis*, 6(2), 21-30.
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories publication bias and psychological science’s aversion to the null. *Perspectives on Psychological Science*, 7(6), 555-561.
- Feynman, R. P. (1974). Cargo Cult Science. *Engineering and Science*, 37(7), 10–13.
- Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of personality and social psychology*, 108(2), 275.
- Gámez, E., Díaz, J. M., & Marrero, H. (2011). The uncertain universality of the Macbeth effect with a Spanish sample. *The Spanish journal of psychology*, 14(01), 156-162.
- Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time*. Technical report, Department of Statistics, Columbia University, New York, NY.
- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean?. *Science translational medicine*, 8(341), 341ps12-341ps12.
- Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic bulletin & review*, 23(1), 74-86.
- Hamermesh, D. S. (2007). Viewpoint: Replication in economics. *Canadian Journal of Economics/Revue canadienne d'économie*, 40(3), 715-733.
- Hendrick, C. (1991). Replication, strict replications, and conceptual replications: Are they important? In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 41–49). Newbury Park: Sage.
- Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge journal of economics*, 38(2), 257-279.
- Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, 66, 81-92.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, 2(8), e124.

- Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6), 645-654.
- 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. doi: 10.1177/0956797611430953.
- Kenrick, D. T., Gutierrez, S. E., & Goldberg, L. L. (1989). Influence of popular erotica on judgments of strangers and mates. *Journal of Experimental Social Psychology*, 25(2), 159-167.
- Kruschke, J. K. (2013). Bayesian Estimation Supersedes the *t*-Test. *Journal of Experimental Psychology: General*, 142(2), 573-603.
- LeBel, E. P., & John, L. (in press). Psychological and institutional obstacles toward more transparent reporting of psychological science. To be published in S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*. New York, NY: John Wiley & Sons.
- LeBel, E. P., Berker, D., Campbell, L., & Loving, T. J. (2016). Falsifiability Is Not Optional. Manuscript currently under review at *Journal of Personality and Social Psychology*.
- LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, 15(4), 371.
- LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org Grassroots Support for Reforming Reporting Standards in Psychology. *Perspectives on Psychological Science*, 8(4), 424-432.
- Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological bulletin*, 70(3p1), 151.
- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives on Psychological Science*, 11(5), 730-749.
- Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, 34(2), 103–115. <http://doi.org/10.1086/288135>
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806–834. <http://doi.org/10.1037/0022-006X.46.4.806>
- Miller, J., & Ulrich, R. (2016). Optimizing Research Payoff. *Perspectives on Psychological Science*, 11(5), 664-691.
- Moery, E., & Calin-Jageman, R. J. (2016). Direct and Conceptual Replications of Eskine (2013) Organic Food Exposure Has Little to No Effect on Moral Judgments and Prosocial Behavior. *Social Psychological and Personality Science*, 7(4), 312-319.

- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 6(7): e1000097. doi:10.1371/journal.pmed1000097
- Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531-536.
- Reinhart, C. M., & Rogoff, K. S. (2010). Growth in a time of debt (digest summary). *American Economic Review*, 100(2), 573-578.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3), 638.
- Rosenthal, R. (1991). Replication in behavioral research. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 1–39). Newbury Park: Sage.
- 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(2), 90.
- Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive psychology*, 22(1), 36-71.
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*, 8(1), 1.
- Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science*, 22: 1359-1366.
- Simmons J., Nelson L. & Simonsohn U. (2012) A 21 Word Solution Dialogue: The Official Newsletter of the Society for Personality and Social Psychology, 26(2), 4-7.
- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9(1), 76-80.
- Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, 26(5), 559–569.
- Simonsohn, U. (2016, March 3). Evaluation replications: 40% full does not equal 60% empty. Retrieved from <http://datacolada.org/47>
- Srivastava, S. (2014, July 1). Some thoughts on replication and falsifiability: Is this a chance to do better? Retrieved from <https://hardsci.wordpress.com/2014/07/01/some-thoughts-on-replication-and-falsifiability-is-this-a-chance-to-do-better/>
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American statistical association*, 54(285), 30-34.
- Taubes, G., & Bond, P. (1993). Bad Science: The Short Life and Very Hard Times of Cold Fusion.
- Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology: Comment on Klein. *Theory & Psychology*, 26(4), 540-548. *Physics Today*, 46(9), 64. <http://doi.org/10.1063/1.2809041>



- Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., ... & STROBE Initiative. (2014). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *International journal of surgery*, *12*(12), 1500-1524.
- Viechtbauer, W. (2015). Package 'metafor'. The Comprehensive R Archive Network. Retrieved from <http://cran.r-project.org/web/packages/metafor/metafor.pdf>
- Vohs, K. D., Mead, N. L., & Goode, M. R. (2006). The psychological consequences of money. *Science*, *314*(5802), 1154-1156.
- Zhong, C. B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*(5792), 1451-1452.

## Appendix A

Prototype UI wireframes of all major pages and views guided by new curation framework:

- home page
- search results page
- evidence collection page (view & edit modes)
- article page
- study page (including data viewer, analytic reproducibility and analytic robustness windows)
- researcher page
- user profile page
- administrator dashboard page

[Currently in development.]