TAKING THE FULL MEASURE:

INTEGRATING REPLICATION INTO RESEARCH PRACTICE

TO ASSESS GENERALIZABILITY

Oleg Urminsky, University of Chicago Booth School of Business
(oleg.urminsky@chicagobooth.edu)

Berkeley Dietvorst, University of Chicago Booth School of Business
(berkeley.dietvorst@chicagobooth.edu)

First Version: May 2022
Current Version: July 2023

*** Please contact authors for updated version before circulating or citing.***

ABSTRACT

In this paper, we review the ways in which replication has been and could be featured in consumer behavior, using JCR as a specific setting. We present a framework for thinking about generalizability of research findings and differentiate various potential benefits that replication can have for understanding variability in consumer research findings. We then define four different types of replications, describe how researchers can use these approaches to produce distinct benefits, and give guidance regarding conducting, interpreting, and the potential contributions of these different types of replications. We conclude with a discussion of various ways in which replication could be more fully integrated into different phases of the scientific research process, taking into account the contribution necessary for publication. In particular, we identify opportunities to incorporate independent replication into original papers, to increase the replication-based contribution in papers that build on prior work and to use systematic replication in conjunction with meta-analysis to synthesize and confirm conclusions from a mature research literature. More fully integrating replication into scientific practice can yield a new equilibrium, in which replication is routine, typically successful, and an accepted prerequisite for establishing an empirical generalization.

Keywords: External Validity, Generalizability, Replication, Conceptual Replication, Direct Replication

"*The ability to systematically replicate research findings is a fundamental feature of the scientific process. Indeed, the idea that observations can be recreated and verified by independent sources is usually seen as a bright line of demarcation that separates science from non-science*." (Dunlap, 1926)

A scientific finding is replicable if new relevant empirical evidence consistently yields a comparable result, consistent with the empirical predictions of a theory. Users of research count on it being replicable, and replicability is a key criterion for transforming data into facts that can serve as a foundation upon which scientific knowledge is built. For research to have real-world impact also relies on replicability: developing reliable advice, applications, and policies requires knowing which findings consistently replicate, and under which conditions.

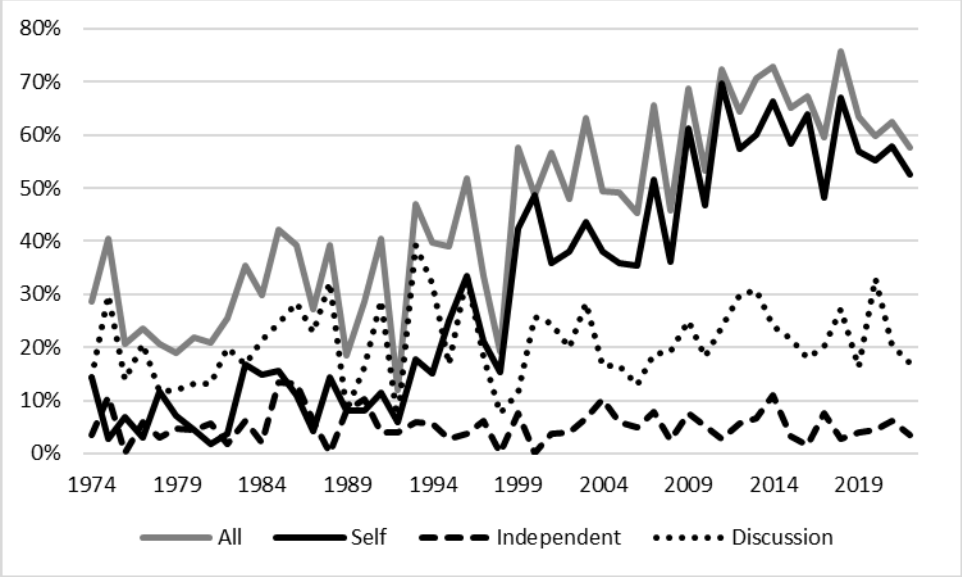
For these reasons, replicability is an ongoing concern for any science that aims to be valid, self-correcting, and practically relevant. Because replication cannot be assumed, a majority of highly cited papers in medicine, for example, are the subject of future replication tests (Ioannidis 2005). After all, the more useful a finding is to its field, in either practical or theoretical terms, the more important it is to be sure of its validity and generalizability, and the more value there should be in testing replication, in some form.

Discussion of replication in consumer behavior research dates back to the year in which the *Journal of Consumer Research* (JCR) was founded (Jacoby 1976) and has become increasingly common since then. In particular, JCR has periodically featured calls for more replication (Helgeson et al 1984; Monroe 1991 & 1992; Hunter 2001; Mick 2003; Rapp 2015) and specific editorial initiatives over the years have attempted to encourage more replication research in JCR (Bettman & Kassarjian 1982; Monroe 1991 & 1992; Mick 2001).

To better understand the nature of replication in consumer research, we conducted a

systematic analysis of all papers published in JCR since its founding.[1] Out of 2664 articles, 1294 (49%) had some discussion of replication. As shown in Figure 1 (solid gray line), mentions of replication in JCR have been increasing dramatically over time (r=.84, p<.001). This increase in mentions of replication over time is primarily due to an increase in within-paper replications of findings. While rare in the early years of JCR, at around 9% of papers, self-replication has now become a norm, with around 54% of papers (and a higher proportion of empirical papers) discussing replications of their own results (r=.92, p<.001). By contrast, reported replication of other papers has remained relatively low over time (mean=5%, r=-.04, p=.80). General discussions of replication, while somewhat more common, have also remained relatively constant (mean=21%, r=.19, p=.19). See the Web Appendix for more details.

Figure 1: Overall and Specific Changes in JCR Replication Mentions



Note: All replication mentions (gray) are broken out into self-replications, primarily within the same paper (black), independent replications of prior findings (dashed), and non-empirical discussion of replication (dotted).

These macro trends illustrate that the concept of replication has long been present and

[1] Additional details and data are publicly available at https://osf.io/my2d3/?view_only=4ea740de89e9458f8885d7459fa4fdcc

is increasing and pervasive in discussions within consumer research and JCR. In fact, as shown in the Web Appendix, JCR has both historically and recently (e.g., in 2022) had a higher rate of mentions of replication than comparable journals in marketing and related fields. In this paper, we attempt to provide more clarity on the use of replication, distinguishing the different methods, benefits, and interpretations of distinct types of replications. First, we describe a framework differentiating potential benefits that replication can have for consumer research. We then define four different types of replications, describe how researchers can use these approaches to produce distinct benefits, and give guidance regarding best practices for, implications of, and the potential contributions of these different types of replications. We conclude by discussing how to evaluate the contribution to research provided by different replication approaches and strategies for integrating replication throughout research practice.

THE NEED FOR REPLICATION

We begin by defining the benefits that replication can provide in general terms. Consider the testable implication of a theory: differences in X should cause (or predict) a change in Y:

$$Y = a + b*X + e \quad (1)$$

This model is empirically testable by measuring X and Y or by manipulating X and observing the resulting differences in Y. The error term *e* means that we cannot expect a change in X to result in the same change in Y every time. Instead, *b* represents an *average estimated* effect corresponding to a specific theoretical prediction or implication, rather than a deterministic effect of X on Y. The convention of treating *e* as *random* error is an assumption of convenience. By definition, *e* actually represents any deviation of actual values of Y from the corresponding predicted value (a + b*X), which includes *all omitted other causes* of Y that

may vary in the world.

In fact, the estimate of *b* from a given dataset may not only vary randomly due to sampling error but is likely to also vary systematically. Thus, we should think of *b* as not only a noisy average, but as a construct that can depend on multiple other factors. For example, we could represent the slope term as:

$$b = b_0 + b_s*Sa + b_P*P + b_T*T + b_C*C + b_{ST}*St + b_M*O + b_{MA}*I + b_E*E + b_R*R + e \quad (2)$$

This formulation represents a hierarchical linear model, in which the estimated coefficient of interest is decomposed into various additive factors that cause (or predict) variation in the coefficient (Yarkoni 2022).

Equation 2 is just one possible representation, specifying common types of variables that the coefficient may vary across, and excluding any interactions. Each b_k may be a vector of coefficients, corresponding to a set of predictor variables of a given type. In this model, Sa denotes the sample of participants, P is for the population the sample was drawn from, T is for time-varying factors, C is for contextual factors, St is for the stimuli, O is for how the outcome (i.e., the dependent variable) was operationalized and measured, I is for how the intervention or predictor (i.e., the independent variable) was implemented, E is for the estimation method, and R is for researcher-specific factors.

The fundamental motivation for replication is that any test will only yield one *specific* estimate of *b*, which will depend on the values of Sa, P, T, C, St, O, I, E, & R. However, consumers of research are often interested in either the *average* effect of X on Y across some representative set of circumstances, or the *expected* effect of X on Y in a particular circumstance. Thus, the need for replication arises from the potential heterogeneity in *b*, which calls into question the validity and generalizability of a specific finding (Lynch 1982; Monroe

1991; Wells 2001; LeBel et al 2017; Fabrigar, Wegener & Petty 2020; Stromland 2021). When a replication varies one or more of the relevant factors that plausibly impact *b*, it can inform what realization of *b* interested parties should expect in a specific setting. As a result, differences across replications that vary relevant factors (as opposed to identical replications) can provide information about how *b* may vary with the factors that differed.

Next, we describe distinct types of replications in terms of which factors (e.g., Sa, P, T, C, St, O, I, E, and/or R) are varied in the replication and which are held constant. We describe what each type of replication can be used for, best practices for conducting each type of replication, and the interpretation of positive and negative results from each type of replication.

COMMON TYPES OF REPLICATIONS

In this section, we describe four distinct and commonly used types of replications: exact, study-design, methodological, and theory-test replications. Our goal is to provide detailed descriptions capturing useful distinctions that may be overlooked in more commonly used terms (e.g., direct vs. conceptual replications). Of course, these four types of replications do not represent a collectively exhaustive list. We focus on these types of replications to provide a conceptual continuum, and to highlight the differences in methods, the unique benefits that each offers, and the different implications of potential results.

The four types of replications are shown in Table 1, ordered from left to right based on the intended similarity to the original study, with exact replications being the most similar and theory-test replications being (potentially) the most different. Further, the replications that appear earlier in the list are largely nested within those that appear later. For example, each exact replication is a study design replication, methodological replication, and theory-test replication; however, a theory-test replication is not necessarily a study design replication. In

Figures 2-5, we provide concrete examples describing how a hypothetical researcher named Avery who is investigating choice overload (whether providing a decision maker with more options can be detrimental to choice; Chernev, Böckenholt, & Goodman, 2015) could employ each type of replication.

Exact Replication

An exact replication involves testing a specific result under exactly the same conditions as the original investigation. Exact replications only differ from the original investigation by the particular sample of participants or data (Sa) included in the study (i.e., all other factors, including the population (P), stimuli (St), and contextual factors (C), are held constant). Because only the sample differs from the focal study, the exact replication tests for variation from the original finding due to sampling error or practices that may capitalize on chance.

One example of an exact replication is cross-validation using a holdout sample: randomly selecting cases from the full dataset that are then used to test the validity of an analysis (typically model-fitting) that has been conducted on the rest of the data. In general, for work using observational data, exact replications are an invaluable tool to address concerns about results being due to random chance (Jacoby 1976), particularly when the methods involve meaningful researcher flexibility. An exact replication can be particularly useful when conducting exploratory research, particularly in circumstances that preclude later replication (*e.g.*, a study conducted in the context of a new product launch).

For experimental work, the only way to create an exact replication is to randomly assign participants to an "original study" and "replication study" at the time of the original study, to hold time-varying factors constant between the original and the replication. While exact replications can address robustness to random chance, other tools may be easier to

implement (e.g., preregistration), and study-design replications (discussed next) can provide similar benefits under stable circumstances that do not preclude later replication.

Table 1: Types of Replications

	Exact replication	**Study-design replication**	**Methodological replication**	**Theory-test replication**
Procedure	Run an identical study in which only the sample varies.	Replicate a particular study at a later date while remaining as true as possible to that original study's design	Replicate the intervention (or other method) from an original study, potentially varying other design choices	Replicate evidence for a particular implication of a theory in a study design that may be different from the original
Benefit	Testing whether original finding is due to chance	Testing robustness of original finding to factors that ostensibly should not change its results	Testing robustness of original study methodology to different research uses	Testing robustness of theory implication to different research designs
Factors that *necessarily differ*	Sample (Sa)	Sample (Sa), Time varying factors (T)	Sample (Sa), Time varying factors (T)	S Sample (Sa), Time varying factors (T), other factor(s)
Factors that *necessarily stay the same*	All except for the Sample (Sa)	Stimuli (St), Intervention (I), Outcome (O), Estimation method (E)	Intervention (I, or other method), Estimation method (E)	None
Factors that *can vary* or *stay the same*	None	Population (P), Contextual factors (C)	Stimuli (St), Outcome (O), Population (P), Contextual factors (C)	Intervention (I), Stimuli (St), Outcome (O), Population (P), Contextual factors (C), Estimation method (E)
Implications if finding *does* replicate	Unlikely that original finding was due to chance. More precise estimate of local *b*.	Evidence that original finding is robust to factors that should not change the results. Lowered estimated heterogeneity in *b*.	Evidence that the methodology produces similar results under different conditions.	Evidence that the tested implication of theory is robust to manipulated factors. Lowered estimate of heterogeneity in *b*.
Implications if finding does *not* replicate	Probable that original finding was due to chance. Estimate of *b* revised downward.	Original finding may not be robust. Overall support for the theory should be reassessed. Estimate of average *b* revised downward, and estimate of heterogeneity in *b* revised upward.	Methodology may not be as robust as originally anticipated. Practitioners and other researchers cannot count on this methodology performing as it did in the original paper.	Evidence that this implication of the theory is not as general as previously believed. Estimate of average *b* revised downward, and estimate of heterogeneity in *b* revised upward.

Exact replications are expected to succeed. When an exact replication does yield a

similar result, we have stronger evidence and a more precise estimate of *b* under the same specific factors. However, findings that are not consistent under exact replication may be spurious (*i.e.*, they may not provide an accurate estimate of *b* even when all possible factors are held constant). A different result in an exact replication requires updating beliefs about the magnitude of *b* from the original study (as in a revised meta-analysis) and could suggest that the procedures involved in generating the original estimate inflated *b* by capitalizing on chance.

Figure 2: Exact Replication Example

Background: Avery tests for choice overload using observational data in which the number of options that consumers faced in a store varies. She decides to employ cross-validation with a holdout sample.
Implementation: Avery randomly assigns observations from the original dataset into "training" and "test" datasets. She runs analyses testing the effect of assortment size on purchase likelihood using the training dataset. Once Avery has settled on a preferred specification in the training dataset, Avery replicates that exact specification in the test dataset.
Benefits: If Avery's preferred specification does not replicate in the test dataset, it reveals that the estimate of *b* she obtained in the training data may not be robust to variation in sample (Sa) and could be due to chance. However, if Avery's preferred specification does replicate in the test dataset, she has obtained a more precise local estimate of *b* in her research context than an equivalent analysis without an exact replication. Both Avery and eventual readers of her work can have more confidence in her estimate of *b* knowing that she conducted an exact replication.
Limitations: Avery's exact replication does not speak to how *b* varies when factors other than sample (e.g., St, O, & I) differ.

Study-design Replication

Study-design replications retain all aspects of the original design judged theoretically relevant. However, such attempts to replicate prior findings are not exact replications, because they do vary one or more factors, typically out of logistical necessity, that are assumed to not matter for the finding (i.e., for the estimate of *b*). For example, the replication may sample from the same population (P; e.g., Mturk participants) but at a later time, or may use a different population that is assumed to be comparable (*e.g.*, undergraduate participants, but at a different institution from the original study), or may involve different contextual factors (C; *e.g.*, an in-person study instead of online). Study-design replications are valuable because they are necessary to obtain a precise estimate of *b* in a particular set of circumstances and speak to the robustness of a particular finding to those factors *assumed to be irrelevant*.

Successful study-design replications provide evidence that a finding replicates in similar circumstances, increasing confidence that the finding can be applied in subsequent research or practical use. However, an unsuccessful study-design replication calls into question whether the original study only provided narrow circumstance-specific evidence for the relevant theory, or was spurious, if actual heterogeneity was low (Coppock, Leeper & Mullinix 2018). This could either mean that seemingly unimportant factors actually represent missing boundary conditions, or that the theory is more generally not supported.

Researchers conducting a study-design replication should clearly specify which factors differ and which remain the same as the original study, because value comes from understanding and interpreting the different conditions generating the new estimate of *b*. For example, a failure to replicate a study with a different population may be a major challenge to the generalizability of an original theory if it did not posit population specific mechanisms, or it could be reasonably expected if the characteristics of that population differ in a way theorized to affect behavior. It is also important to test any manipulation checks or other measures of necessary pre-conditions, to distinguish between a failure to replicate the conditions needed to test a theory versus a successful test of theory that failed to replicate the predicted result.

For study-design replications, the two factors that necessarily differ from the original study are the sample (Sa) and time varying factors (T, representing factors that may have changed in the world since the original study took place). Theories in consumer behavior are generally not intended to be specific to a particular draw of participants or a particular point in time, as a theory that only applied to a single sample or single point in time would be of little value, either theoretically or practically. Thus, consumers of research will typically expect results of a study to be robust to variation in these factors. Failures to replicate under different

time-varying factors or with a different sample should typically be interpreted as a serious challenge to the findings from the original study.

However, there can be rare exceptions. It is important for the researchers conducting a replication to contemplate whether any differences in time varying factors, sample characteristics, or other factors provide a theoretical or logistical reason for the result to differ. For example, if recent world events have changed the way that participants interpret the stimuli of a study or if a stimulus was chosen for specific perceptions at the time of the study (i.e., a newly released product), a later study-design replication may no longer provide a valid test of the relevant theory. Ideally, the original research would have clearly specified such necessary conditions and included empirical tests that could then be confirmed prior to a replication.

Figure 3: Study-design Replication Example

Background: Avery is building on a prior lab study that manipulated the number of restaurant menu options that participants had in a hypothetical choice and found evidence for a new choice overload outcome – an increased likelihood of asking for advice. She is now replicating that study with an online sample.
Implementation: Avery takes the study design from the lab study and makes the minimal changes necessary to run it with an online sample. She considers whether stimuli or other factors need to change to provide an equivalent test of the theory, given the potential changes in time varying factors (T) or population (P), and only makes changes that are necessary. For example, if a dish in the original choice set is only known regionally, she may decide to update the choice set with an otherwise similar option, ideally using pre-tests to check the similarity of the updated stimuli.
Benefits: If the results of the study design replication are not consistent with the previous lab study, Avery learns that the results of the original study may not be robust or may vary meaningfully with factors that she had assumed were unimportant. Avery should reassess all of the evidence she has collected to update both her estimate of average *b* and her beliefs about heterogeneity in *b*. Avery may also consider follow-up testing to see whether any of the factors that differed between studies moderate choice overload. If Avery finds evidence for choice overload consistent with her original lab study, she has gained a more precise local estimate of *b*, and evidence that the original finding is robust to the presumed-irrelevant changes in factors that differed between studies. Both Avery and readers should have more confidence in the finding.
Limitations: Avery's study-design replication may speak to how choice overload varies with sample (Sa), time varying factors (T), and population (P), but does not speak to how *b* varies with fundamentally different study designs. Future work is needed to learn how choice overload may be different with different stimuli (St), outcomes (O), independent variables (I), and etcetera.

Methodological Replication

Methodological replications test the robustness of the methods in an original study (e.g., the intervention used) to changes in other factors. Unlike study replications, methodological replications may include major changes to the study, like updated stimuli (St)

or different outcome measures (O). In particular, methodological replications may focus more on validation measures, such as manipulation checks or scale validation tests, than on the original outcome. The results of methodological replications speak to whether an intervention, scale, or other methodology performs as it was proposed to in the original research.

As is the case for study replications, researchers conducting methodological replications should clearly describe what factors they changed from the original study, because the contribution of the replication is to test generalizability of the method to those changes. Successful methodological replications provide some evidence that the method can reliably be used in a new research or applied setting (i.e., that the interpretation of *b* when using the method is relatively robust to the particular changes in Sa, T, St, Me, P, C, and/or E). Failures to replicate suggest that the effects of an intervention or interpretation of a measurement method may be more variable than originally believed (i.e., the meaning of *b* may depend on factors specific to the original study).

Methodological replications have clear practical implications for researchers and practitioners who are considering adopting a method from past work. Original research that introduces a new intervention to the literature, for example, makes a potential methodological contribution by creating a tool that others can employ. However, the extent of this contribution depends on whether the intervention consistently produces similar effects in different circumstances. Thus, methodological replications are important for determining whether other researchers can count on a method working as intended.

Figure 4: Methodological Replication Example

Background: Avery is testing decision task difficulty as a moderator of choice overload. She runs a methodological replication to see if a manipulation of font from the fluency literature effectively manipulates the difficulty of her decision task.
Implementation: Avery runs a study manipulating the font used in the decision task with the goal of understanding how this manipulation will perform in her context. She uses the stimuli (St) and population (P) from her planned study, and adds outcome (O) measures to determine if the manipulation has the desired effect on decision task difficulty.

Benefits: If Avery does not replicate the manipulation of font in her context, both Avery and readers of her work learn that the effect of this manipulation (*b*) may vary with factors like stimuli (St) or contextual factors (C). Future researchers and practitioners should consider testing the efficacy of this manipulation before employing it in their context of interest. If Avery finds that the font manipulation has similar effects on decision task difficulty as it had in the prior contexts, she has learned that it may be a viable manipulation for her research. Both Avery and consumers of her work learn that the effect of this manipulation (*b*) may be robust to the stimuli (St), population (P), contextual factors (C) that differ in her study.
Limitations: Avery's methodological replication will not provide evidence regarding whether the original fluency studies replicate because she has changed multiple factors including the stimuli (St), outcome (O), and contextual factors (C).

Theory-test Replication

Theory-test replications implement a *new* test of the *same* theoretical implication or prediction (i.e., the same *b*) as the original study did, potentially changing multiple or even all factors (Sa, P, T, C, St, O, I, E, and/or R) relative to the original. Importantly, a theory-test replication is distinct from a study that tests a *different* theoretical implication or prediction from the original study, which, while informative about the theory, is not a replication, because it estimates an entirely different *b*.

The benefit of theory-test replications is that testing the same prediction of a theory using a completely new methodology provides substantial variation in the potential influences on *b*, such that a successful replication may provide strong evidence of generalizability (Blanchard et al., 2022). However, the results should be interpreted with caution. The more differences there are from the original study, the greater the potential concern that what seems like a replication may in fact instead be a similar-seeming effect that is due to different causes.

Furthermore, when many factors vary simultaneously, a failure to replicate does not identify specific limits on generalizability, as the researcher will not have strong evidence that the difference in results is attributable to any particular manipulated factor. Thus, while successful theory testing replications may increase one's confidence that the relevant theory will hold in untested circumstances, failures often suggest that more work is needed to understand exactly what factors the relevant theory is contingent on, to refine relevant theory.

For this reason, a series of theory-testing replications, each varying one factor that would be relevant for drawing conclusions about generalizability, may be most informative.

Figure 5: Theory-Test Replication Example

Background: Avery tests for choice overload with a previously unused outcome measure (O). Specifically, she will test whether purchases from larger assortments are more likely to be returned.
Implementation: Avery adopts a previously used study design, and makes the outcome measure (O) returns of the selected product instead of the original outcome measure (choice deferral). Avery is careful not to manipulate other important features of the study design like the stimuli (St) or manipulation (I) so that she can reasonably infer that any differences in results may be attributable to the change in outcome measure (O).
Benefits: If Avery's test does not replicate, she has gained evidence that choice overload may not be as general across outcome measures (O) as previously assumed. Both Avery and consumers of her work increase their expectations about the extent to which *b* is likely to vary across outcome measures (O). If Avery finds evidence in line with choice overload, she has gained evidence that choice overload is more general across outcome measures (O) than the past empirical evidence has shown. Both Avery and consumers of her work may lower their estimate of the heterogeneity in *b* when outcome measures (O) are changed.
Limitations: Avery's theory testing replication does not provide the strongest possible evidence regarding whether the previously used study design replicates because one important factor, the outcome measure (O), differed.

We can relate these four types of replications to some commonly used terms. Exact replications are sometimes called out-of-sample tests. Study-design replications are sometimes called direct replications (although that term is not well-defined) or, incorrectly in our view, exact replications. Methodological replications may be referred to as method validation. Finally, theory-test replications have been referred to as conceptual replications (e.g., Lynch et al 2015), although that term has also sometimes been used to refer to testing different theoretical implications, which we do not consider to be replications. We have developed the present model-based taxonomy of replication types to reduce the potential for confusion, and to create terminology that we hope will increase the precision and constructiveness of conversations about replication, by being clearer about the mapping between the methods and claims in an original paper and a particular replication.

SELF-REPLICATIONS VERSUS INDEPENDENT REPLICATIONS

Finally, we note that researcher-specific factors (R) add another dimension to the distinctions between replications that we have discussed so far. For example, study-design,

methodological, and theory-testing replications can be conducted by the same author(s) that produced the original work (self-replications) or by independent researchers (independent replications) or may range somewhere in between.

Assessing the sensitivity of *b* to researcher-specific factors is particularly relevant to debates about replication. In the scientific ideal, the results of a replication are not specific to the particular set of researchers conducting it (i.e., $b_R = 0$). Specifically, if all relevant research design, implementation, and analysis factors have been accounted for, researcher characteristics should have no remaining effect. In practice, however, there could be a systematic researcher-specific effect (i.e, b_R may be nonzero) due to omitted factors that differ by researcher. For example, there may unspecified boundary conditions that only the original authors are aware of, idiosyncratic research practices that are specific to a particular researcher (e.g., use of a particular attention check, or a particular data cleaning practice), or even research practices that inflate false-positive rates (Simmons et al 2011, John et al 2012). "Open science" practices such as sufficiently detailed and publicly posted pre-registrations and shared stimuli and analysis code are beneficial in this regard, by enabling the field to identify the specific conditions under which a given result was obtained, which would otherwise constitute unknown researcher-specific effects.

The potential for researcher-specific variation in *b* constitutes a limitation of relying primarily on self-replication. Because self-replications keep R fixed, they cannot identify researcher effects. In addition, researchers advancing a particular theoretical viewpoint (and perhaps trying to convince a skeptical review team) may, intentionally or not, skew their choice of research design factors (e.g., population (P), stimuli (St), etcetera) towards those under which the effect of interest is more likely to be observed. Sometimes, as when the stated

research goal is an "existence proof," this may be directly discussed in the paper, while in other cases the non-representativeness may not be acknowledged.

Successful independent replications suggest that the methodology used in the original research is sufficiently complete and robust to be employed by different researchers or practitioners (i.e., when researcher-specific factors vary). As a result, marketing managers should be more comfortable employing practices from research that has been replicated independently. Likewise, other researchers may be more confident in developing projects that build on research that has already been replicated independently.

As discussed earlier, our review finds that the rate of JCR papers presenting specifically *independent* replication is low: historically, only 5% of papers, compared to the 36% of papers that discuss replicating their own results. This data suggests that, intentionally or not, the field of consumer behavior has addressed validity concerns involving replication primarily through self-replication, typically within the original paper. In fact, it is noteworthy that the current rate of independent replication is lower in JCR than in some comparable journals (e.g., 3% of replication mentions in JCR in 2022, compared to 40% in the *Journal of Experimental Psychology: General*; see Web Appendix).

Current standards for publication increasingly dictate the "replication and systematic search for range and limits" stage of research development (Wells 1993) taking place within an original paper. Presenting replication evidence alongside the proposed theoretical advance can be a faster, more standardized and more efficient alternative to relying on independent replication. However, even extensive self-replication in original papers cannot address the potential for researcher-specific effects. Evidence from multi-site pre-registered replications of highly cited findings in the behavioral sciences (e.g., Klein et al 2018) suggest that researcher

effects can exist. Thus, for all its benefits, self-replication alone seems unlikely to fully address issues of robustness and generalizability.

TOWARDS AN INTEGRATION OF REPLICATION INTO RESEARCH PRACTICE

In order for diverse replications that can more fully assess generalizability to become more prevalent in consumer research, the field needs to grapple with the practical need to evaluate the contribution of a replication. Advocates of more replications, including specifically in JCR, have proposed criteria for assessing the contribution of a project centered around replication, such as the importance of the phenomenon, the novelty and influence of the specific paper being replicated, and the reasons for suspecting lack of generalizability (Monroe 1991 & 1992; Mick 2001; Lynch et al 2015). Below, we build on these discussions to propose specific ways for authors to incorporate different types of replications into their original research, throughout the scientific process, while considering the need for original research to meet the bar in terms of contribution.

Within-paper replication.

The current prevalence of self-replication (typically within a single paper) in consumer behavior research should be recognized as informative for assessing generalizability and a clear indication of the field's concern with validity. The potential for self-replication to increase generalizability could be dramatically increased with detailed pre-registration and by consistently reporting (e.g., in an online repository) all relevant studies collected, along with well-documented data and analysis code, crucially including any non-replications. For replication within an original paper to serve as the primary mode of assessing generalizability, the comprehensiveness and credibility of self-replication during the review process would require the same scrutiny typically applied to questions of internal validity (e.g., confounds in

experimental design) and construct/argument validity (e.g., theoretical development).

To specifically address the potential for researcher-specific factors, papers could go beyond self-replication. In much the same way that researchers often rely on independent coders unaware of the hypotheses, papers could include a study-design replication conducted by an independent replicator, which could also be a useful tool for addressing reviewer skepticism. We are currently developing a pilot program to conduct independent replications for authors who wish to include such a study in their paper.

Independent Replication Commentaries.

Some independent study-design replications of previously published results will be small in scope, empirically and/or theoretically, relative to the contribution of the original paper being replicated and may therefore not meet the contribution standards for a full research paper. Self-publication (e.g., blog posts) can miss the benefits of peer review, potentially makes the replications more difficult to find (particularly in the long term) and may be more easily dismissed by the field. The field may be best served by publishing such replications using an alternative format, such as brief commentaries or letters, as currently done in *Psychological Science* (Bauer 2021) and previously in JCR (Bettman & Kassarjian 1982; Mick 2001).

However, such replications should not be automatically assumed to make less of a contribution than original work. Replication-based papers that are both rigorously conducted and sufficiently ambitious, theoretically and empirically, to be informative on an important question can provide as much or more contribution than a more typical empirical paper, whether providing "good" or "bad" news. This may involve broadening the focus from a single finding or paper to a reassessment of the full literature that bears on the broader motivating question (*e.g.*, by conducting an "empirical audit" on that topic, O'Donnell et al 2022).

Replication During Subsequent Theory Development.

In many cases, a more direct path for replication-related work to make a sufficient contribution will be to go beyond retesting a specific finding to conducting theory-test or methodological replications that have broader theoretical or practical contributions. A paper could "replicate and extend/moderate" prior findings, not only confirming but providing a more detailed understanding. Many papers already do this, often focusing on the extension or moderation as the key contribution but sometimes failing to highlight the replication aspect and to convey what has been learned about generalizability. Alternatively, a "non-replication and explanation" would demonstrate a failure to replicate, identify reasons for the differing results, often factors previously assumed to be irrelevant, and update the theory accordingly.

Like any cumulative research, these approaches risk being perceived as "incremental," particularly if the original paper already provided substantial evidence regarding generalizability. Authors need to make the case for the importance of the theory being refined and for the informativeness of the new evidence. This may be accomplished either by challenging the conclusions or by reducing substantial uncertainty from the original paper.

Replication While Testing Other Hypotheses.

Research on a novel hypothesis often contains studies that are, in part, study-design or theory-test replications of the prior work that the paper builds on. Similarly, studies that use previously developed methods in new ways may constitute methodological replications, if designed to validate the method. Such studies could be designed with comparability to prior research and the resulting replication-based secondary contribution in mind. The replication contribution would also be bolstered by explicitly mentioning and sharing all replication-relevant data, with relevant citations, regardless of results. A brief summary of the replication-

based implications, both positive and negative, of the paper's findings, in the GD or appendix, would inform the field about the generality of prior research. Far more independent replication is likely being conducted than what is apparent from reading published papers.

However, it is important to acknowledge potential disincentives to authors of doing this, particularly the risk of alienating authors whose work may not have replicated. To generate a meaningfully *representative* dataset of independent replications from original research (vs. only successful replications) requires that reviewers and other stakeholders recognize the value of and promote the practice of full reporting of replication-relevant results.

Research Synthesis Through Meta-Analytic Replication.

Given that the benefits of theory-testing replication are primarily in assessing generalizability, the greatest need for systematic replication of these types may, in fact, come late in the research process. Science is typically cumulative, with general conclusions derived from comparison and synthesis of all the findings in a literature, in the form of qualitative (theoretical) or quantitative (meta-analytic) reviews. However, even a mature research literature may not contain sufficient, and sufficiently varied, replications to reach conclusions about empirical generalizations.

This can be remedied by making replication a key element of conducting research synthesis. Once a field is sufficiently developed for a meta-analysis on a particular question, the criteria for replication research to make a sufficient contribution are likely to have been met (i.e., the phenomenon has been deemed important, the prior work is novel and influential, and conflicting results or moderators identified in the prior work make the case that replication is needed.) Pre-registered, sufficiently powered and independently conducted confirmatory replications of the meta-analytic conclusions would provide exactly the kind of unbiased

estimates that typically cannot be assumed from traditional meta-analyses alone (Kvarven et al 2020). Meta-analytic replication has been used to form conclusions about ego-depletion (Vohs et al 2021) and pain of paying (Bechler et al 2022).

GENERAL DISCUSSION

Ideally, replication of one's own and others' findings would be a standard and unremarkable aspect of conducting research. In practice, frictions and disincentives can limit the amount and types of replication findings that make it into the literature, distorting the evidentiary basis upon which consumers of research rely. When independent replication is rare, non-replication may be seen as severe criticism (as opposed to just more data) or even as a tacit accusation of fraud, raising the stakes for reporting incidental non-replication findings.

When a field lacks consensus regarding the benefits of replication in general and how to assess the contribution of a specific replication, attempts to publish replications may be especially uncertain, costly, and frustrating, influencing what researchers with scarce resources prioritize (Romero, 2018). In this light, meta-scientific findings of non-replication (e.g., Klein et al 2018) and perceptions of replication studies as "policing" are a symptom that a field has insufficiently fostered a culture of self-correcting science, with distinct negative consequences.

In this paper, we have attempted to promote the practice of replication by reducing some of these frictions for authors. We attempted to formally define the benefits of replication in a way that integrates the conduct and assessment of replication into the basic practice of theoretical and practical research. We then drew distinctions between common types of replications that involve different goals and correspondingly different methods, to provide a common vocabulary and shared guidelines for researchers. We describe the different ways in which incorporating replication can bolster the contribution of a specific research paper and

outline specific types of papers that can be developed in line with each path to contribution. In sum, we discuss how to use replication to contribute to theory-driven empirical research, by testing generalizability and thereby providing the evidence for theory confirmation or disconfirmation, theory extension, or theory integration and expansion.

Given the costs of producing empirical papers, it is important for replications to be integrated throughout existing research practices, rather than constituting additional costs. Our framework suggests how authors can execute and communicate the study-design, theory-test, and methodological, replications that are already being conducted as part of original research. We are hopeful that authors can use independent replication to address some reviewer concerns more effectively than running additional original studies. We give suggestions outlining how authors of replication-based papers may achieve and communicate sufficient contribution. Finally, while meta-analytic replication would involve collecting additional data, this may be effective as a follow-up paper to a high-contribution meta-analysis paper, or as an approach that can complete a meta-analysis paper that would otherwise have insufficient contribution.

Replicability enables consumer behavior research to build on prior findings by identifying substantiated theories and validated methods, and to identify promising new questions and provide reliable guidance to consumers, practitioners and policy makers. The dramatic increase in self-replication, seen in the scope and content of JCR articles over the past 50 years, indicates a concern for replicability. More fully incorporating a variety of replication approaches (including independent and meta-analytic confirmatory replication) throughout the research process would further improve the validity of consumer research. Integrating replication throughout scientific practice can yield a new equilibrium, in which replication is routine, typically successful, and a prerequisite for establishing an empirical generalization.

REFERENCES

Bauer, Patricia J. (2021), "A New Option for Scientific Exchange and an Alternative to the Commentary Format," *Psychological Science*, 32(9), 1343-45.

Bechler, Chris, Rhia Catapano, Szu-chi Huang, and Oleg Urminsky (2022), "The Pain of Paying Effect Revisited: Replicating the Past, Moderating the Present, and Informing the Future," working paper

Bettman, John and Harold Kassarjian (1982) "Editorial," *Journal of Consumer Research*

Blanchard, Simon J., Jacob Goldenberg, Koen Pauwels, and David A. Schweidel (2022), "Promoting Data Richness in Consumer Research: How to Develop and Evaluate Articles with Multiple Data Sources," *Journal of Consumer Research*, 49(2), 359-72.

Chernev, Alexander, Ulf Böckenholt, and Joseph Goodman (2015), "Choice Overload: A Conceptual Review and Meta-Analysis," Journal of Consumer Psychology, 25(2), 333-58.

Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. (2018) "Generalizability of heterogeneous treatment effect estimates across samples." *Proceedings of the National Academy of Sciences* 115, no. 49: 12441-46.

Dunlap, K (1926), "The Experimental Methods of Psychology," in *Psychologies of 1925*, ed. C. Murchison, Clark University Press, 331–53.

Fabrigar, Leandre R., Duane T. Wegener, and Richard E. Petty (2020), "A Validity-Based Framework for Understanding Replication in Psychology," *Personality and Social Psychology Review*, 24(4), 316-44.

Helgeson, James G., E. Alan Kluge, John Mager, and Cheri Taylor (1984), "Trends in Consumer Behavior Literature: A Content Analysis," *Journal of Consumer Research*, 10(4), 449-54.

Hunter, John E. (2001), "The Desperate Need for Replications," *Journal of Consumer Research*, 28(1), 149-58.

Ioannidis, John P. (2005), "Why Most Published Research Findings Are False," *PLoS Medicine*, 2(8).

Jacoby, Jacob (1976), "Consumer Psychology: An Octennium," *Annual Review of Psychology*, 27(1), 331-58.

John, Leslie K., George Loewenstein, and Drazen Prelec (2012), "Measuring the Prevalence Of Questionable Research Practices with Incentives for Truth Telling," *Psychological Science*, 23(5), 524-32.

Klein, Richard A., Michelangelo Vianello, Fred Hasselman, …, and Brian A. Nosek (2018), "Many Labs 2: Investigating Variation in Replicability across Samples and Settings," *Advances in Methods and Practices in Psychological Science*, 1(4), 443-90.

Kvarven, Amanda, Eirik Strømland, and Magnus Johannesson (2020), "Comparing Meta-Analyses and Preregistered Multiple-Laboratory Replication Projects," *Nature Human Behaviour*, 4(4), 423-34.

LeBel, Etienne P., Derek Berger, Lorne Campbell, and Timothy J. Loving (2017), "Falsifiability Is Not Optional," *Journal of Personality and Social Psychology*, 113(2), 254-61.

Lynch Jr, John G. (1982), "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research*, 9(3), 225-39.

Lynch Jr, John G., Eric T. Bradlow, Joel C. Huber, and Donald R. Lehmann (2015), "Reflections on the Replication Corner: In Praise of Conceptual Replications," *International Journal of Research in Marketing*, 32(4), 333-42.

Mick, David (2001), "From the Editor," *Journal of Consumer Research*

_____(2003), "Editorial," *Journal of Consumer Research*

Monroe, Kent (1991), "Editorial: On Replications in Consumer Research, Part I," *Journal of Consumer Research*

_____(1992), "Editorial: On Replications in Consumer Research, Part II," *Journal of Consumer Research*

O'Donnell, Michael, Amelia S. Dev, Stephen Antonoplis, …, and Leif D. Nelson (2021), "Empirical Audit and Review and an Assessment of Evidentiary Value in Research on the Psychological Consequences of Scarcity," *Proceedings of the National Academy of Sciences*, 118(44).

Raman, Kalyan (1994) "Inductive Inference and Replications: A Bayesian Perspective," *Journal of Consumer Research*, 20(4), 633-43.

Rapp, Justine M., and Ronald Paul Hill (2015), "'Lordy, Lordy, Look Who's 40!' The Journal of Consumer Research Reaches a Milestone," *Journal of Consumer Research*, 42(1), 19-29.

Romero, Felipe (2018), "Who Should Do Replication Labor?" *Advances in Methods and Practices in Psychological Science*, 1(4), 516-37.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22(11), 1359-66.

Strømland, Eirik (2021), "Making Our 'Meta-Hypotheses' Clear: Heterogeneity and the Role of Direct Replications in Science," *European Journal for Philosophy of Science*, 11(2), 1-19.

Vohs, Kathleen, Brandon Schmeichel, Sophie Lohmann, …, and Dolores Albarracín (2021), "A Multisite Preregistered Paradigmatic Test of the Ego-Depletion Effect," *Psychological Science*, 32(10), 1566-81.

Wells, William D. (1993), "Discovery-Oriented Consumer Research," *Journal of Consumer Research*, 19(4), 489-504.

_____(2001), "The Perils of N= 1," *Journal of Consumer Research*, 28(3), 494-98.

Yarkoni, Tal (2022), "The Generalizability Crisis," *Behavioral and Brain Sciences*, 45.

WEB APPENDIX

Reviewing and coding prior research.

We used EBSCO Host to code for different kinds of discussion of replication in prior research. Our primary analysis focused on coding articles in the *Journal of Consumer Research* (JCR) from the journal's founding in 1974 until 2022. The coding was conducted by the first author and a research assistant. The first author first reviewed a set of recent articles to identify terms used to discuss replications, and to develop a set of categories capturing different ways in which replication has been discussed in the literature. The coding was then conducted by the first author and the research assistant. The full coding data files are publicly available at https://osf.io/my2d3/?view_only=4ea740de89e9458f8885d7459fa4fdcc.

We first searched for articles with any of the partial terms: replic* or nonreplic* or unreplic* or reanaly* in any of the text, as shown in Figure A1 below, resulting in 1468 articles identified. Articles with any of the search terms were then coded for the type of discussion of replication contained in the article, for all of the following categories that applied to each paper:

1. Replication within same study (i.e., a result in one condition or subsample replicates a result in another)
2. Replicates (part or all of) a prior study in the same paper
3. Replicates/reanalyzes a study from another paper with at least one author in common
4. Replicates/reanalyzes at least one study in another paper with all different authors
5. Directly recommends replication of own results
6. Discussion of replication either in the literature review, the discussion of results or the GD
7. Editorial (i.e., discussing need for or prevalence of replication)
8. Replication is mentioned in the title of cited paper(s)

9. Directly recommends replication of others' results
10. Irrelevant meanings (e.g., use of replicates to mean copies, repeated measures, elements of experimental design, bootstrap samples, etc…)
11. Unclear

In the analysis of JCR papers, we also accounted for the fact that 25 of the papers discussing replication appear twice in the EBSCO database (due to being featured a second time in "Research Collections"; listed below). Total papers for the years 2013 and 2014 were coded from the JCR website due to this issue: https://academic-oup-com.proxy.uchicago.edu/jcr/issue. The results for the full time period are presented in Figure A2.

We also conducted two analyses to compare across journals. First, we conducted the same search on EBSCO for JCR and 9 other journals, both for the time period 1974-2022 (or since the journal began, if after 1974) and for 2022 alone:

Academy of Management Journal (AMJ), 1974-2022

American Economic Review (AER), 1974-2022

Journal of Experimental Psychology: General (JEP:G), 1974-2022

Journal of Marketing (JM), 1974-2022

Journal of Marketing Research (JMR), 1974-2022

Journal of Personality and Social Psychology (JPSP), 1974-2022

Management Science (MGSCI), 1974-2022

Marketing Science (MKSCI), 1982-2022

Psychological Science (PS), 1990-2022

For this analysis, we did not exclude the duplicate JCR paper because we were not able to perform the same check for duplication for the other journals. We also intended to conduct the same search in *Cognition* and *Journal of Consumer Psychology*, but we were not able to access the information for these journals, which are not carried in EBSCO.

We show a comparison of the proportion of articles discussing replication across articles, both overall and in 2022, in Figure A3. This analysis shows that JCR has historically had a higher rate of mentions of replication than the other journals and has continued to do so in 2022.

Next, we computed an index, as the ratio of the percentage of articles that mentioned replication in 2022 relative to the percentage of articles that mentioned replication in its full history, for each journal. We plot the results in Figure A4. Six journals, including JCR, had indices above 1, indicating that there were more mentions of replication in 2022 than previously. By contrast, four journals had indices lower than one, indicating fewer mentions of replication in 2022 than in the past.

Finally, to compare the types of mentions across journals, we coded articles published in 2022 in two additional journals, the *Journal of Marketing Research* and *Journal of Experimental Psychology: General*. We coded all 57 articles that mentioned replication in JMR in 2022, and the chronologically first 60 out of the 220 articles that mentioned replication in JEP:G in 2022. The results are shown in Figure A5. It is notable that independent replications were substantially higher in JEP:G than in the two marketing journals in 2022.

An important caveat to all of these analyses should be noted: we analyze the frequency and content of replication *mentions*, not the frequency and content of replications. This is because we rely on the author(s) reporting and characterization of replication in their papers. We suspect the rates of replication (successful or not) of other authors' findings are substantially higher, because many papers building on prior research do conduct theory-test or even study-design replications but may fail to present them as

such. We focus on mentions (i.e., instead of coding papers for our judgment of whether they replicate a prior finding) because we are attempting to characterize the replication information that a reader of the literature would have easy access to.

Figure A1: EBSCOhost Search terms

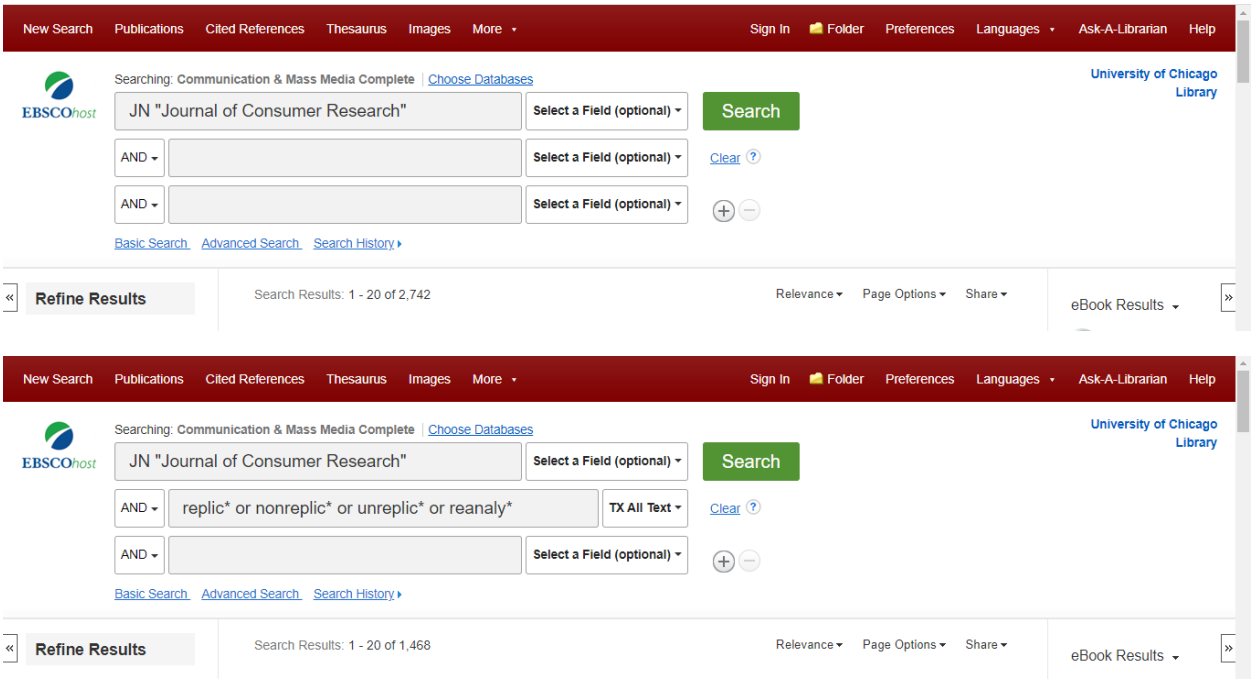
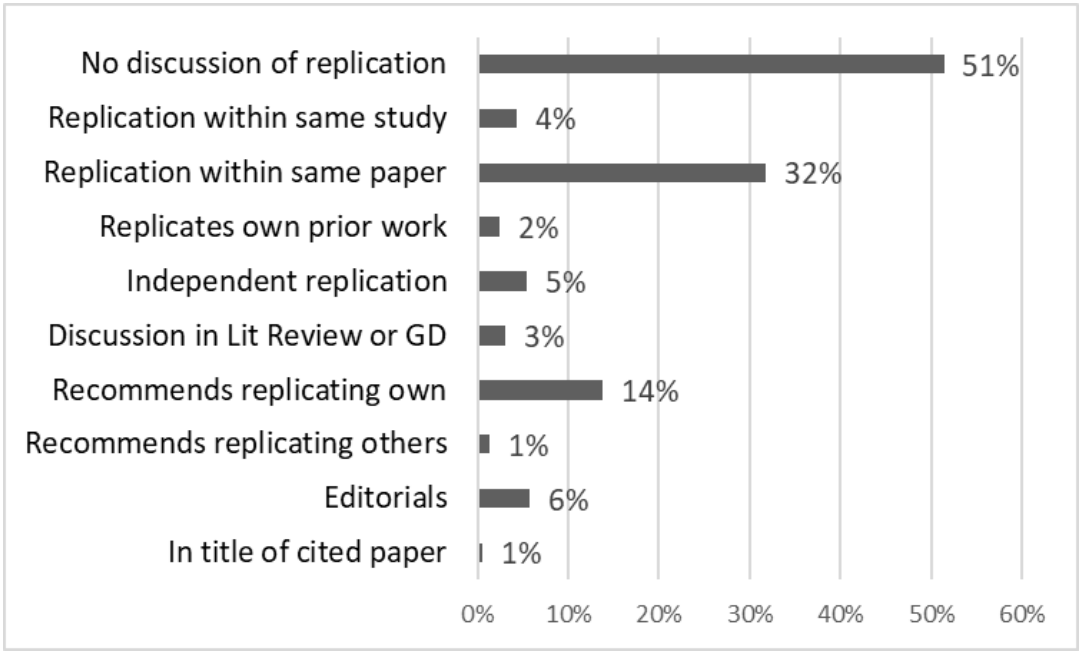
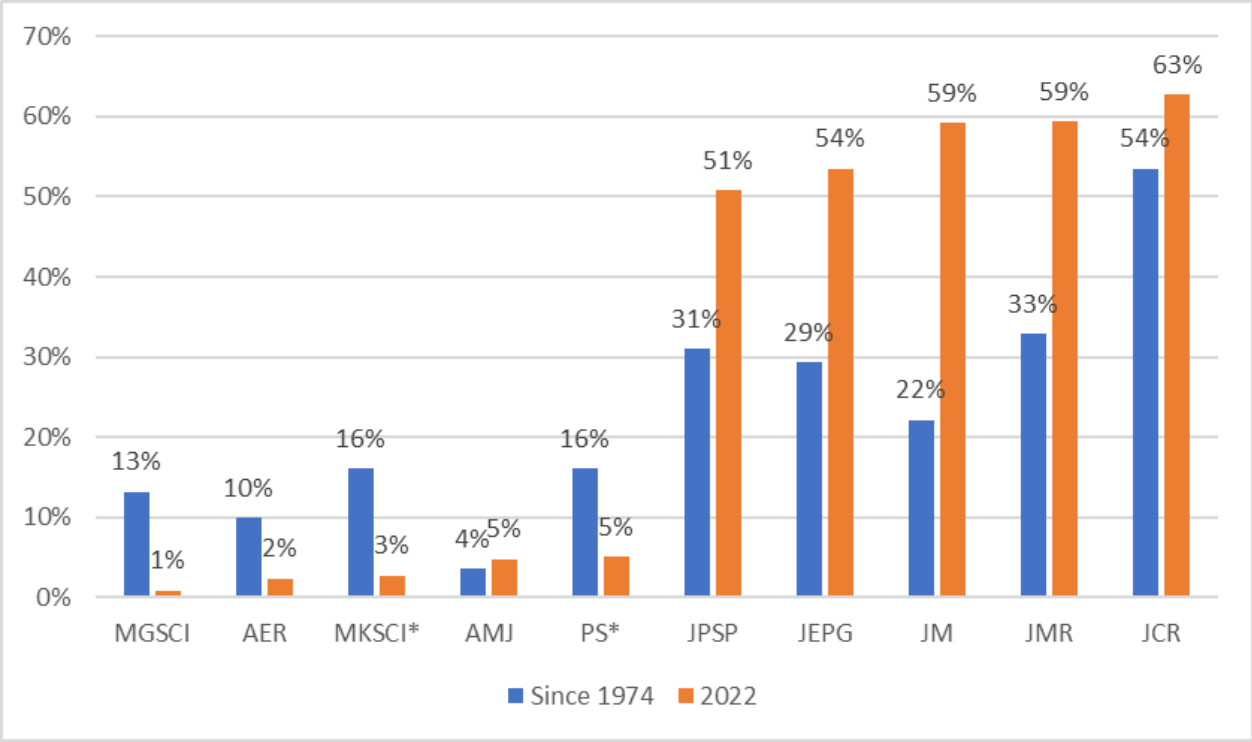


Figure A2: Different types of mentions of replications in JCR, 1974-2022



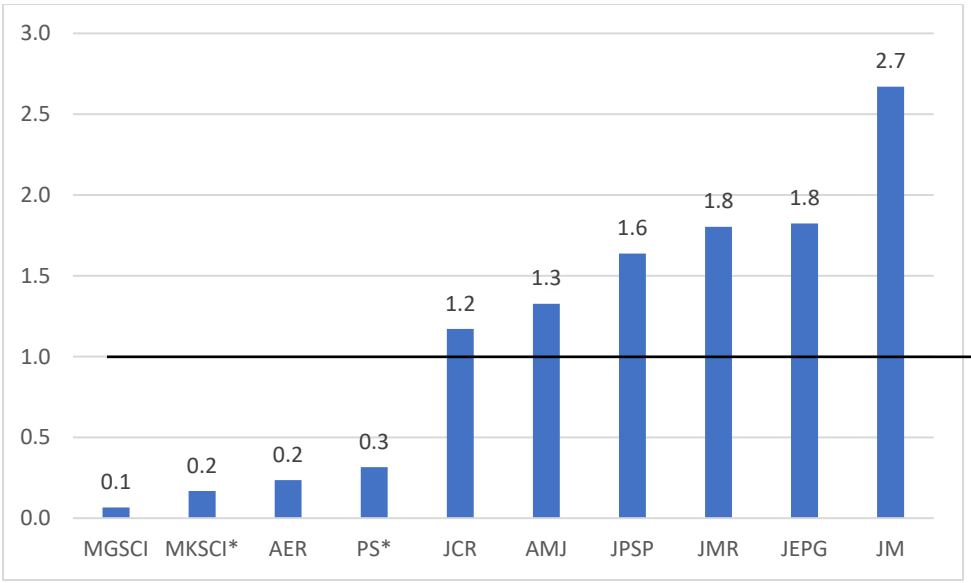
Note: The numbers do not sum to 100% because articles may qualify for multiple codes

Figure A3: Comparing frequency of replication discussions across journals



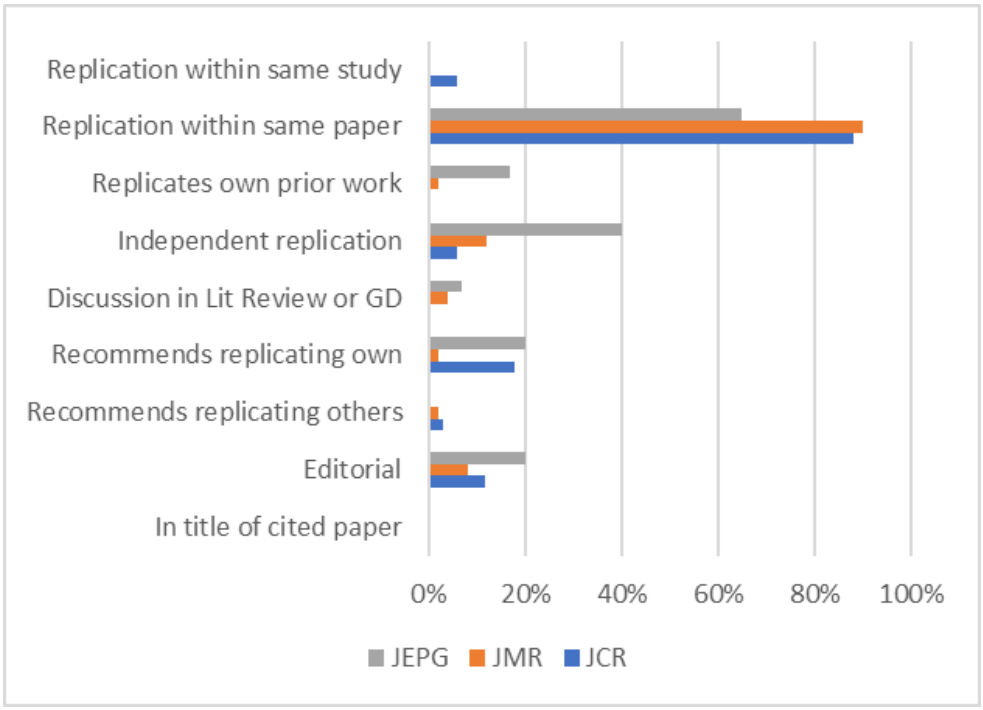
*Note: *Psychological Science* began in 1990 and *Marketing Science* began in 1982.

Figure A4: Comparison of replication mentions in 2022 to full history, by journal



Note: Psychological Science began in 1990 and Marketing Science began in 1982. Index is calculated as % of articles that mentioned replication in 2022 divided by the % of articles that mentioned replication in full history.

Figure A5: Comparison of types of mentions of replication across three journals in 2022



Note: Percent of all papers mentioning replication

Duplicate JCR Articles in 2013 and 2014 “Research Collections”:

"I Don't" versus "I Can't": When Empowered Refusal Motivates Goal-Directed Behavior.

An Arousal Regulation Explanation of Mood Effects on Consumer Choice.

An Interpretive Frame Model of Identity-Dependent Learning: The Moderating Role of Content-State Association.

Are Close Friends the Enemy? Online Social Networks, Self-Esteem, and Self-Control.

Dissociative versus Associative Responses to Social Identity Threat: The Role of Consumer Self-Construal.

Financial Deprivation Prompts Consumers to Seek Scarce Goods.

Getting Ahead of the Joneses: When Equality Increases Conspicuous Consumption among Bottom-Tier Consumers.

Guiltless Gluttony: The Asymmetric Effect of Size Labels on Size Perceptions and Consumption.

How and Why 1 Year Differs from 365 Days: A Conversational Logic Analysis of Inferences from the Granularity of Quantitative Expressions.

How to Make a 29% Increase Look Bigger: The Unit Effect in Option Comparisons.

Identifiable but Not Identical: Combining Social Identity and Uniqueness Motives in Choice.

I'll Have What She's Having: Effects of Social Influence and Body Type on the Food Choices of Others.

Illusionary Progress in Loyalty Programs: Magnitudes, Reward Distances, and Step-Size Ambiguity.

Lenses of the Heart: How Actors' and Observers' Perspectives Influence Emotional Experiences.

Nostalgia: The Gift That Keeps on Giving.

Self-Identity and Consumer Behavior.

Show Me the Honey! Effects of Social Exclusion on Financial Risk-Taking.

Small Sounds, Big Deals: Phonetic Symbolism Effects in Pricing.

The Impact of Product Name on Dieters' and Nondieters' Food Evaluations and Consumption.

The Signature Effect: Signing Influences Consumption-Related Behavior by Priming Self-Identity.

Too Much of a Good Thing: The Benefits of Implementation Intentions Depend on the Number of Goals.

Toward a Theory of Status Consumption in Less Industrialized Countries.

When Healthy Food Makes You Hungry.

When Imitation Doesn't Flatter: The Role of Consumer Distinctiveness in Responses to Mimicry.

Years, Months, and Days versus 1, 12, and 365: The Influence of Units versus Numbers.

Current Pilot Plan for Independent Replication Initiative:

1. Researchers volunteer to conduct replications (referred to as "replicators" hereafter).
    - Replicators will be recruited from multiple institutions, to help deal with conflict of interest issues.
    - Volunteers need to first establish feasibility in terms of their institution's IRB policies. One such institution has been identified thus far.
2. Authors file a request in the online portal and upload their materials to their own OSF folder:
    - Approved IRB application from own institution
    - Pre-registration
    - Stimuli, ideally a Qualtrics .qsf file
    - Analysis plan or code, ideally in R
    - Draft or template language for reporting results

    The Authors can pre-specify contingencies such as criteria for the study to be a valid test,
    e.g., manipulation checks or minimal levels on outcome variables.
3. Randomly assigned non-conflicted Replicator reviews request for fit with criteria
    - Only appropriate for confirmatory analyses where all conclusions can be pre-specified
    - Initial criteria: Low participant risk, no deception, online implementation
    - Longer-term goal: Ability to run in-person lab studies
4. If approved, Replicator files IRB with their own institution.
5. Authors pay for participant compensation fees and online platform costs.
    - Authors can request funding for these costs based on having limited research resources. Initially, funding for four such projects will be available.
6. After IRB approval, Replicator runs study, runs code and updates draft language with results. Replicator posts all files in a separate OSF folder managed by the replication initiative and completes an online form documenting the conclusion of the project.
7. Authors include study writeup in their own paper.

8. At Authors' request, read-only access link to project-specific OSF folder is provided, to be included in paper or shared with reviewers.
9. 3 years after the date of the study, the OSF folder including the original request and author identity is automatically made public, and the project is listed on the web page.