aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2015, Vol. 10(3) 390–399

DOI: 10.1177/1745691614564879
pps.sagepub.com
SAGE

Replicating Studies in Which Samples of Participants Respond to Samples of Stimuli

Jacob Westfall[1], Charles M. Judd[1], and David A. Kenny[2]
[1]University of Colorado Boulder and [2]University of Connecticut

Abstract
In a direct replication, the typical goal is to reproduce a prior experimental result with a new but comparable sample of participants in a high-powered replication study. Often in psychology, the research to be replicated involves a sample of participants responding to a sample of stimuli. In replicating such studies, we argue that the same criteria should be used in sampling stimuli as are used in sampling participants. Namely, a new but comparable sample of stimuli should be used to ensure that the original results are not due to idiosyncrasies of the original stimulus sample, and the stimulus sample must often be enlarged to ensure high statistical power. In support of the latter point, we discuss the fact that in experiments involving samples of stimuli, statistical power typically does not approach 1 as the number of participants goes to infinity. As an example of the importance of sampling new stimuli, we discuss the bygone literature on the *risky shift phenomenon*, which was almost entirely based on a single stimulus sample that was later discovered to be highly unrepresentative. We discuss the use of both resampled and expanded stimulus sets, that is, stimulus samples that include the original stimuli plus new stimuli.

Keywords
replication, stimulus sampling, statistical power

There has been a recent surge of interest in conducting replication research in psychology. This interest has arisen because of storied failures to replicate highly cited effects (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012), growing concerns about undisclosed flexibility in the conduct of psychological research (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), and occasional but attention-grabbing reports of data fabrication (Enserink, 2012). As a result, a variety of recent papers have stressed the importance of routinely conducting replications of published results and publishing these replication attempts (Francis, 2012; IJzerman, Brandt, & Van Wolferen, 2013; Ioannidis, 2012; Koole & Lakens, 2012; Nosek, Spies, & Motyl, 2012; Simons, 2014). At least three journals (*Journal of Experimental Psychology: General, Journal of Personality and Social Psychology*, and *Perspectives on Psychological Science*) now regularly devote space to articles that attempt to replicate prior studies, and other journals (e.g., *Social Psychology;* Nosek & Lakens, 2014) have released special issues consisting entirely of replication studies. Moreover, there is an ongoing large-scale collaborative effort that seeks to empirically estimate the actual replicability of findings from top psychology journals (Open Science Collaboration, 2012).

Although the general enthusiasm for conducting replication studies is at a record high, several important questions remain concerning the precise details of how researchers should think about, conduct, and evaluate replication attempts. Providing widely agreed-upon answers to these questions is an increasingly urgent project for psychological researchers; some contributions to this project include Brandt et al. (2014), Schmidt (2009), Simonsohn (2013), and Verhagen and Wagenmakers (2014). In this article, we aim to further clarify some lingering methodological questions surrounding replication by discussing one aspect of the replication enterprise that

Corresponding Author:
Jacob Westfall, University of Colorado–Psychology, UCB 345, University of Colorado Boulder, CO 80309
E-mail: jake.westfall@colorado.edu

we believe has been insufficiently scrutinized: the role of stimulus sampling and the handling of stimulus materials. In doing so, we challenge some of the recent guidelines and recommendations that have been offered.

The Nature and Purpose of Replication

In discussions of replication, it is important to distinguish among three sorts of replication studies: *conceptual, direct*, and *exact replications* (Brandt et al., 2014; Schmidt, 2009). In a conceptual replication of a study, the outcome variables, experimental manipulations, participant populations, and so on might all differ from those in the original study. Nevertheless, the original finding is considered to be conceptually replicated if it can be convincingly argued that the same theoretical constructs thought to account for the results of the original study also account for the results of the replication study (Stroebe & Strack, 2014). Conceptual replications are thus "replications" in the sense that they establish the reproducibility of theoretical interpretations. By contrast, a direct replication seeks more specifically to reproduce the methods of the original study. Direct replications are intended to verify that particular experimental results cannot be attributed to sampling error, multiple comparison problems, reporting biases, and so on. Finally, an exact replication implies that all of the conditions of the original study are implemented again in the replication: the same participants, experimenters, locations, materials, and so on. Exact replications are impossible—some things must necessarily differ from the original study (minimally, the dates on which the studies were conducted), otherwise the replication is literally the same study and not a distinct study at all—but the notion of an exact replication is nevertheless instructive as a thought experiment.

Using the terms defined above, it is direct replications that have seen such an impressive surge of interest in recent years, and it is direct replications that we are concerned with in this paper. Ultimately, the goal of direct replication is to confirm that a particular experimental procedure can reliably produce a particular empirical result (Schmidt, 2009).

To help make our discussion of direct replication more concrete, consider a simple psychological experiment in which participants are randomly assigned to one of two groups: an experimental group in which the participants undergo some theoretically motivated treatment procedure, or a control group in which the participants undergo a comparable but theoretically inert placebo procedure. Assume that, in this experiment, the responses on some relevant behavioral measure were systematically different in the experimental condition than they were in the control condition. Now suppose that an independent laboratory wishes to directly replicate this finding before beginning to examine possible extensions. How should the direct replication proceed?

Although there may be some disagreement regarding specific methodological details of the direct replication, there are two general considerations that appear to be widely agreed upon. The first is that although the replication should closely follow the experimental procedure of the original study, and the sample of participants in the replication study should be drawn from the same or a comparable population as the original sample, the new sample of participants should not be the exact same participants who were recruited for the original study.

The reason for changing the sample is that one of the main purposes in conducting a replication study is typically to rule out the possibility that the original results can be attributed to sampling error in the recruitment of participants. If the same sample of participants were to be used, and the original finding replicated, it is doubtful whether we really would have gained anything from the replication other than increased confidence that the treatment procedure leads to detectable group differences for those particular participants.

The second consideration is that the replication study should have a high level of statistical power. It is, of course, important for any study to have high statistical power—a fact of which there is a growing awareness in psychology (Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013; Ioannidis, 2008; Schimmack, 2012). High power is particularly important for replications, so as to avoid the needless confusion and potential controversy that could follow if the findings of the original study were not successfully replicated, despite perhaps being true. Indeed, a recent policy statement on research practices from the Society of Personality and Social Psychology has recommended that "In the case of replication research, hallmarks of high quality include adequate power (and the more the better, perhaps suggesting a benchmark of 0.90 or 0.95 for adequate power for single replication studies rather than the conventional 0.80)" (Funder et al., 2014, p. 9).

Sampling Error Versus Generalizability

We thus see two considerations as essential for direct replications: using new samples from the same or similar populations and ensuring high statistical power. Before moving on, we want to clearly distinguish these two considerations from considerations of the generalizability of a finding. The question of generalizability concerns whether a theoretical result can be produced under a range of different conditions, such as with different participant populations, in different settings, with different outcome variables, and so on. Establishing the generalizability of a finding is thus the purview of conceptual

replication. Although establishing the generalizability of research findings is undoubtedly important work, it is not the focus of this article (for opposing viewpoints on the value of conceptual replications, see Pashler & Harris, 2012; Stroebe & Strack, 2014). Ultimately, the two considerations outlined above are concerned with accounting for sampling error due to participants. We seek new participants (drawn from the same or a similar population to the original participants) to guard against the possibility that the original sample of participants happened to be unusual. And we desire high statistical power in order to minimize the impact of sampling error in the replication study.

We do not mean to imply that these two considerations are the *only* functions of direct replication. Other important functions of direct replication can be to investigate whether the original results depended on theoretically extraneous factors such as the laboratory in which the studies were conducted, to determine whether the original results were likely the result of reporting biases or questionable research practices, and in some rare cases, to help rule out the possibility of data fraud. Our point is that although there may or may not be these additional motivations behind a particular replication attempt, it would presumably be of interest in *any* direct replication to examine the role of sampling error in the original study and to ensure adequate statistical power in light of that sampling error.

Replicating Studies That Employed Samples of Participants and Samples of Stimuli

We now consider a more complicated class of experiments that is very common in psychology: experiments involving samples of participants responding to samples of stimulus materials (Judd, Westfall, & Kenny, 2012; Wells & Windschitl, 1999; Westfall, Kenny, & Judd, 2014). Common examples include memory studies in which participants memorize lists of words that are drawn from a larger corpus of words, studies of social cognition in which participants make judgments about sets of faces or read vignettes about hypothetical persons, and studies of emotion in which participants are exposed to photographs or film clips of emotion-provoking scenes. A feature shared by all of these experiments is that the stimulus materials can be best understood as a sample of stimuli drawn from some theoretical stimulus population of interest. Referring to the stimulus set as a "sample" does not imply that the stimuli are selected haphazardly: Indeed, often great attention and care is taken in selecting the particular stimuli that are ultimately used in a study. Nevertheless, in these experiments there are, in principle, other possible stimuli that could have served the experimenter's purposes just as well as those that were in fact selected, and it is in this sense that the stimulus set is a sample. Using statistical terminology from the literature on analysis of variance, the stimuli are properly understood as a *random factor* (as opposed to a *fixed factor*) in these experiments. Thus, these experiments are more complicated than the simple two-independent-groups example just discussed in that they involve multiple random factors: random participants as before and random stimuli as well.

To the extent that there is variance in these experiments that is attributable to stimuli[1] (e.g., some words are more memorable than others or some faces are judged differently than others), then analyses of the resulting data are biased unless stimuli are in fact treated as a random factor (Clark, 1973; Coleman, 1964; Kenny, 1985; Wells & Windschitl, 1999). Recent advances in statistical methods and software permit such analyses using *mixed models* with reasonable facility (Baayen, Davidson, & Bates, 2008; Judd et al., 2012).

For a direct replication of a study involving sampled stimuli, we can presume the same two broad considerations mentioned previously still apply in this case. First, we wish to replicate the original findings using the same procedures and populations as in the original study but with new samples in order to account for sampling error in the original studies. Second, steps should be taken to guarantee that the replication study should have high statistical power. Unfortunately, virtually all of the published sets of guidelines and recommendations concerning the conduct of direct replications fail to ensure that either of these considerations will be met for studies involving samples of stimuli. We explain the reasons for this in detail below.

Using new samples

The first consideration of using new samples in a replication study is that replications of studies that originally employed both a sample of participants and a sample of stimuli should ideally include not only a new sample of participants (drawn from a population of participants comparable to the original study), but also a new sample of stimuli (drawn from a population of stimuli comparable to the original study). If all of the direct replications of a finding involve the same sample of stimuli used in the original study, and if all of these replications successfully find the same results, this would indeed greatly increase our confidence that these results hold for this particular set of stimuli. However, these replications would not increase our confidence that we would find the same results using other comparable samples of stimuli. To increase our confidence, we have to actually use other samples of stimuli. Just as it is desirable to use a

new sample of participants to help guard against the possibility that the original study results were solely attributable to an unusual participant sample, it would also be desirable to use a new sample of stimulus materials when replicating the study. In the next major section, we discuss in some detail an historical example from social psychology involving research on the *risky shift phenomenon*, in which researchers were misled for years by an unusual stimulus sample that was subsequently used by most replicating researchers.

The foregoing argument about using new stimulus samples to account for sampling error is fairly straightforward and intuitive. However, it leads us to a policy that is in direct contradiction to recent recommendations about direct replications. For example, the Reproducibility Project—the large-scale collaborative effort to estimate replicability mentioned at the beginning of this article—instructs its participating researchers that "Replications must […] use the original materials, if they are available" (Open Science Collaboration, 2012, p. 658). Moreover, a comprehensive set of guidelines on the "replication recipe" for conducting a convincing direct replication lists the second ingredient of the recipe as "Following as exactly as possible the methods of the original study (including participant recruitment, instructions, stimuli, measures, procedures, and analyses)" (Brandt et al., 2014, p. 218).

We think that these recommendations should be revised. We do acknowledge that, in many cases, it would indeed be desirable for a replication study to use mostly the same materials as the original study. But whether this is true for the stimuli to which participants respond depends on whether those stimuli are best understood as having been sampled from some theoretical population of stimuli. On the one hand, if the conclusions of the study are truly intended to apply only to the particular set of stimuli that were used, then clearly a direction replication should use the same stimuli. For example, in a study of mental arithmetic focusing on participants' ability to multiply two positive single-digit integers, there are only 81 possible products to be studied, and hence it is entirely possible that the stimulus set would fully exhaust the theoretical population of interest. On the other hand, if there are, in principle, other similar stimuli (e.g., other words or other faces) that could have served the purposes of the research just as well as those that happened to have been used, and if in fact there is variability in responses attributable to stimuli, then researchers conducting a direct replication would do better to seek new but comparable stimuli (see also Monin & Oppenheimer, 2014). In our experience, studies in which the stimulus set fully exhausts the population of interest are rare indeed.

We do note that random stimuli and random participants are not completely parallel in all respects. One notable difference is that participants are often changed by their experience in a study (e.g., through practice effects, increased knowledge of the experimental conditions, and so on), whereas stimuli are not. Thus it can often be problematic to reuse the same participants in a replication study, whereas reusing stimuli is not as problematic. Although this difference is certainly true, we do not think that it is a convincing justification for reusing stimuli. For instance, imagine that these concerns did not apply to the participants; that is, imagine that the participants were not changed by their experience in the study, so that we could in principle reuse the same set of participants as many times as we wished. Would most psychologists therefore view it as perfectly acceptable to test the same set of participants over and over again? We do not think so. Instead it would still be desirable to draw a new sample of participants in each experiment in order to account for sampling error, as discussed extensively above.

Ensuring high statistical power

Now we turn to the second major consideration, specifically, that the replication study should have high statistical power. In situations where participants are the only random factor, ensuring high statistical power is relatively simple in principle. If we recruit enough participants, we would eventually have high statistical power. When both participants and stimuli are random, the issue of statistical power is more complicated because power is a joint function of both of the sample sizes—that is, the number of participants and also the number of stimuli (Westfall et al., 2014). Surprisingly, when stimuli are treated as an additional random factor crossed with participants, recruiting additional participants (holding constant the number of stimuli) will not always lead to sufficient statistical power. Instead, as the number of participants approaches infinity, power converges to a maximum attainable power value that is almost always lower than 1—in fact, the maximum attainable power value can potentially be quite small.

Consider a hypothetical experiment in which a sample of participants all respond to a sample of stimuli and these stimuli are nested in one of two experimental conditions. For example, in social psychology, we might have participants making judgments about photographs of either White or African American males; in cognitive psychology, we might have participants recalling words that are either short (small number of syllables) or long (large number of syllables) from a word list that they previously studied (Baddeley, Thomson, & Buchanan, 1975). As shown in Westfall et al. (2014), the highest possible level of power in such an experiment (i.e., if an infinite number of participants were recruited, but the stimulus sample remained unchanged) is a function of

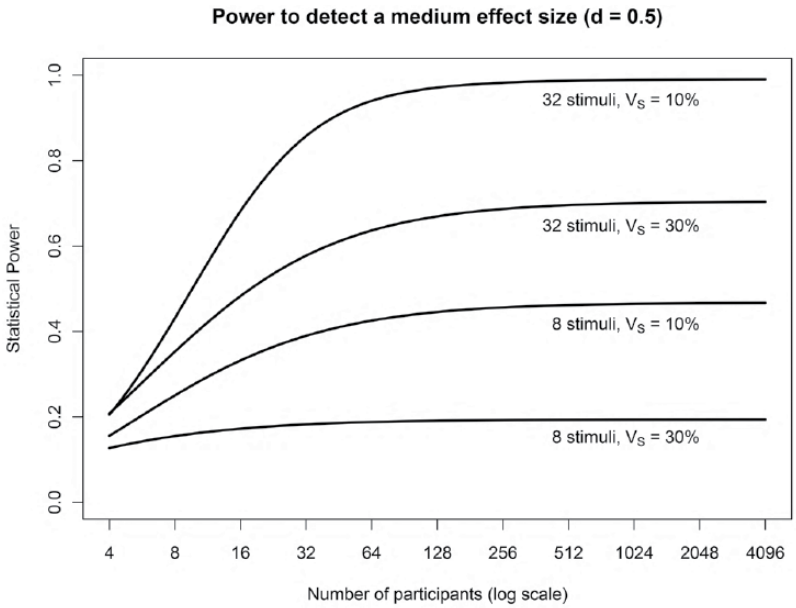

Fig. 1. Plot of statistical power as a function of the total number of participants for the stimuli-within-condition design. The term V_S is the proportion of the total variation in the data that is due to stimulus variance. Power effectively reaches its asymptote at around 200 participants or so, and statistical power at this asymptote can be quite small. The power values plotted here rely on some reasonable assumptions about the full set of variance components in the experiment (Westfall, Kenny, & Judd, 2014). Note that these assumptions only affect the rate at which the power values converge to their asymptotes; they do not affect the maximum power values, which depend only on the effect size, number of stimuli, and V_S.

the effect size, the number of stimuli, and the degree of stimulus variability, and that highest possible power level can be much less than 1.

In Figure 1, we plot the power to detect a medium effect size (d = 0.5) for such a study as a function of the number of participants, the total number of stimuli, and the relative variability of the stimuli. There are several important things to notice in the figure. First, even with very large numbers of participants, statistical power generally does not asymptote at 1, as it does when the only random factor is the participants. Rather, as the number of participants increases, statistical power approaches a maximum value that can be much less than 1. Second, the maximum attainable power value in a study depends critically on the stimulus sample, specifically, on the number of stimuli and on the proportion of the total variance due to stimuli (denoted V_S). For instance, with a sample of 32 stimuli (16 in each condition) that exhibit a substantial degree of variability (V_S = 30%), the maximum attainable power is about .70. But with a much smaller sample of only 8 stimuli (4 in each condition), even if the stimuli are much less variable (V_S = 10%), the maximum attainable power is less than .50. Third, the increases in statistical power gained by augmenting the number of participants show diminishing returns at relatively low values: Increasing the number of participants alone beyond 64 or so does little to increase power, assuming stimuli are a random factor and they account for at least some variance in the data.

One implication of this analysis is that if a direct replication of a study involving stimulus sampling follows the traditional advice about increasing the number of participants but uses the exact same stimulus set, then it would often be *theoretically impossible* for the power of the replication study to be much higher than that of the original study. If the stimulus sample in the original study was relatively small, and especially if the sample was highly variable, then the maximum attainable power of the replication study is likely to be correspondingly small, no matter how many participants the replication study recruits. In other words, when stimuli are sampled, the two apparently reasonable recommendations of using the same stimulus set and achieving high statistical power are often in direct contradiction. Therefore, in addition to our previous argument that a direct replication study should ideally involve a new sample of stimulus materials, we also strongly suggest that it would often be advisable to augment the size of this stimulus sample in order to ensure that the replication study has adequately high statistical power. If it makes sense that replication studies

should employ a greater number of participants than the original study for statistical power purposes, then it makes sense that replication studies should employ a greater number of stimuli as well.

The Risky Shift and the Choice Dilemma Questionnaire: A Cautionary Tale

There is a highly relevant example in the history of psychology in which the failure to appreciate the importance of stimulus sampling in the context of replication systematically misled researchers for about 10 years. This is the case of the risky shift phenomenon, the investigation of which became an area of very active research in social psychology in the 1960s (Cartwright, 1971, 1973; Myers & Lamm, 1976; Pruitt, 1971). Our purpose in this section is not to provide a detailed review of the early research on the risky shift. Instead, we examine the early risky shift literature as a case study in which theoretical progress was unnecessarily impeded by multiple generations of replication studies nearly all relying on the same stimulus sample, which turned out to be unrepresentative, in a nonobvious but important way, of the domain it was intended to represent.

The risky shift phenomenon refers to the idea that, following a group discussion in which the members collectively provide advice to someone considering a risky decision in some hypothetical context, the members of the group tend to favor the risky decision more than they had before the group discussion. The classic risky shift paradigm involved a sample of participants responding, individually at first, to a series of 12 items known as the Choice Dilemma Questionnaire (CDQ; Kogan & Wallach, 1964; Stoner, 1961). For example, one of the items described a Mr. A, an electrical engineer with a secure job but with a modest salary, who has been offered a potentially lucrative job at a newly founded company with a highly uncertain future. The participant's task is to indicate the lowest probability of the company proving financially sound that they would consider acceptable for Mr. A to change jobs. After participants completed the CDQ, they convened in small groups to discuss all of the CDQ items as a group and render a group decision for each item. A risky shift occurred when the group decisions indicated a greater willingness to endorse the risky decisions, on average, than the individual group members had indicated in their personal responses to the same items beforehand.

The first demonstration of the risky shift, using the CDQ, was published as a Master's thesis by Stoner (1961). The finding was immediately studied by social psychologists around the world. Cartwright (1973) determined that, about 10 years after Stoner (1961), the risky shift literature had seen 196 papers by 187 investigators from eight countries. This rapid interest occurred for several reasons. The idea of a risky shift was contrary both to the conventional wisdom of the time and to classical social psychological theory. It addressed an important problem with direct relevance to many real-world situations. And it could be easily replicated. Most of the replication studies following in the wake of Stoner (1961) employed the CDQ as their stimulus set, and they generally had no trouble obtaining the basic risky shift result.

Today we know that there is no risky shift. That is to say, decisions following group discussion are not always or even usually more risky than the same decisions rendered by individuals. Fifteen years after the first published experiment on the risky shift phenomenon, Myers and Lamm (1976) remarked

> "It is now widely recognized that the designation *risky shift* was a misnomer [...] The risky shift label continued to guide experimentation long after it was well established that shift to greater caution could be reliably demonstrated on certain choice-dilemma items" (p. 603).

Cartwright (1971) wrote

> "It is now clear that the items contained in the original CDQ are in no sense a representative sample of the universe of all possible items. Instruments similar to the CDQ could readily be constructed whose scores would display risky shifts, cautious ones, or none at all" (p. 368).

When greater care was taken to examine the risky shift hypothesis using more than just the original 12 CDQ items, it was found that the notion of a generally risky shift was, at best, an overly simplified view of what happens to individual attitudes following group discussion (Pruitt, 1971).

Looking back on these events, an optimistic view is that they represent a successful demonstration of the much-storied self-correcting nature of science. A less rosy view is that, considering the amount of time and resources that were spent, researchers should have been quicker to identify that the basic premise on which the entire literature was based and which it ultimately sought to explain—that group discussion leads people to favor more risky decisions—could not be reproduced with comparable items other than those found in the CDQ. If researchers today were to adopt a policy of varying rather than duplicating the randomly sampled stimulus materials from original studies, perhaps in the future it would not take as much time and effort to uncover such basic problems as it did in the case of the risky shift.

Table 1. Schematic for an Example Replication Study Using an Expanded Stimulus Set

	Original stimulus block				New stimulus block				New stimulus block				New stimulus block			
Participants	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	A	A	B	B	-	-	-	-	-	-	-	-	-	-	-	-
2	A	A	B	B	-	-	-	-	-	-	-	-	-	-	-	-
3	A	A	B	B	-	-	-	-	-	-	-	-	-	-	-	-
4	-	-	-	-	A	A	B	B	-	-	-	-	-	-	-	-
5	-	-	-	-	A	A	B	B	-	-	-	-	-	-	-	-
6	-	-	-	-	A	A	B	B	-	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-	-	A	A	B	B	-	-	-	-
8	-	-	-	-	-	-	-	-	A	A	B	B	-	-	-	-
9	-	-	-	-	-	-	-	-	A	A	B	B	-	-	-	-
10	-	-	-	-	-	-	-	-	-	-	-	-	A	A	B	B
11	-	-	-	-	-	-	-	-	-	-	-	-	A	A	B	B
12	-	-	-	-	-	-	-	-	-	-	-	-	A	A	B	B

Note: This particular design involves a single fixed factor (Condition, with two levels, Condition A and Condition B) and two crossed random factors (Participants and Stimuli). In this design, Participants are crossed with Condition, while Stimuli are nested within Condition. If a cell in the table below contains an A, it means that this participant responded to this stimulus only under Condition A. If a cell contains a B, it means that this participant responded to this stimulus only under Condition B. If a cell contains a dash, it means that this participant never responded to this stimulus. In the design shown here, each participant is randomly assigned to respond to one of four stimulus blocks, each block containing four stimuli. The original stimulus block (Stimuli 1 to 4) contains all and only the original set of stimuli. Each new stimulus block (Stimuli 5 to 8, 9 to 12, and 13 to 16) contains a resampled set of stimuli.

Using Resampled Stimulus Sets Versus Expanded Stimulus Sets

Up to this point, we have argued that replication studies should routinely employ entirely resampled stimulus sets of adequate size to ensure high power. But we also like an alternative strategy: the use of what we call "expanded" stimulus sets. An expanded stimulus set is one that includes the original stimuli plus new stimuli drawn from the same or a comparable stimulus population.

The use of expanded stimulus sets has much to recommend it. First, there may be cases when one wishes to show that some demonstrated effect really is confined to the stimuli originally sampled (Fiedler, 2011). In these cases, one might wish to conduct a replication study using an expanded stimulus set, with the goal of showing that the stimuli do matter and that successful replication can be accomplished only with the original stimulus sample. Second, in the case of an unsuccessful replication, the use of an expanded stimulus set can help clarify whether the failure to replicate is likely attributable to the new participant sample, the new stimulus sample, or other factors.

In Table 1, we show an example of how a replication study employing an expanded stimulus set might be designed. As before, in this hypothetical study we consider a design in which participants are crossed with the experimental conditions whereas the stimuli are nested within the experimental conditions; for example, participants completing a memory task in which their memory for previously studied concrete and abstract nouns is assessed (Gorman, 1961). One notable feature of this replication study is the use of what Westfall et al. (2014) refer to as a stimuli-within-block design. In these designs, the full stimulus sample is divided into a smaller number of comparable lists or blocks, and each participant is randomly assigned to receive only one of these blocks. In this particular implementation, the first stimulus block corresponds to the set of stimuli used in the original study, whereas the other stimulus blocks are composed of new stimuli drawn from a similar population. One advantage of this design is that the number of responses made by each participant can be held constant and equal to what it was in the original study, while the size of the stimulus sample is still augmented considerably, which can substantially benefit the statistical power of the replication study (as discussed by Westfall et al., 2014). However, expanded stimulus sets need not necessarily be used in the context of stimuli-within-block designs. Often it will be sufficient simply to have the participants in the replication study respond to a greater number of stimuli than in the original study.

Conclusion

In psychology, the presumption is that participants matter: Different participants give different responses. Therefore, if an effect is to be demonstrated, one needs some sufficient number of participants so that variability attributable to them does not mask the effect. The

variability attributable to participants is thus treated as random error and an effect of interest is only judged as meriting attention if it results in variance that is large relative to the variance associated with participants.

In many experimental paradigms, stimuli also matter. Indeed, the fact that different responses are given to different stimuli constitutes the essence of psychology. But frequently, individual stimuli to which responses are given are simply instances that are sampled from categories of such stimuli, and it is the categories that are the real focus of experimental investigation. In such cases, presuming that there is variance attributable to stimuli, then that variance ought also to be treated as error variance, just as is variance attributable to participants (Baayen et al., 2008; Clark, 1973; Coleman, 1964; Judd et al., 2012; Kenny, 1985; Wells & Windschitl, 1999).

Attempts to replicate experimental results naturally assume that replications should be carried out with new samples of participants to account for sampling error and that these new samples should be sufficiently large to guarantee high power. Our argument is that when stimuli are also random samples of possible stimuli that could be used, and when in fact different stimuli elicit different responses, then the strategies of replication research that apply to participants ought also to apply to stimuli. That is, generally new samples of stimuli should be used to demonstrate that the previous effects are not simply due to the original stimuli that were sampled. In addition, a sufficient number of stimuli should be used to ensure high statistical power. These recommendations contradict many extant suggestions for how replication research ought to be conducted.

Would we advocate that replication research never use the same stimuli as the original study? Clearly new stimuli are not always required. First, if the stimuli exhaust all or nearly all of the possible stimuli (e.g., the lists of English consonants and vowels, or of single-digit numbers), then arguably there are no other possible samples that might be used. Thus, what we have to say here applies only to studies in which the stimuli used are only partial samples of stimuli that might have been used to instantiate some class of stimuli of theoretical interest. Second, if stimuli do not *matter* (i.e., there is no variability in responses attributable to them), then replication research need not use new stimulus samples. But again, we think that stimuli typically do matter, and of course it is an empirical issue whether or not they do: The absence of stimulus variance cannot be established in an a priori manner. Third, there are sometimes good reasons to use expanded stimulus samples, using the original sample and an additional new sample, as discussed in the previous section.

If stimuli matter, then statistical power is affected by the variability that they induce in the responses. Accordingly, in replicating some effect of theoretical interest, statistical power is a joint function of the number of participants sampled and the number of stimuli sampled. It is also important to note that, to the extent that stimuli matter, there are upper limits to statistical power as a function of increasing the numbers of participants and that these limits are very likely to be considerably less than 1. Guidance in choosing sample sizes for both participants and stimuli to maximize power, as a function of the proportion of variance expected from each, is available in Westfall et al. (2014).

Research involving stimulus samples is ubiquitous in psychology. Just as replication researchers attend to their participant samples, so should they attend to their stimulus samples. And such attention should not always mean blindly assuming that the same stimuli ought to be used in a replication study as in the original study. If stimuli matter and if they are sampled, then like participants, they ought to be resampled, and resampled in sufficient numbers to guarantee high power.

Declaration of Conflicting Interests

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

Note

1. We note that there are really two distinct sources of stimulus variation that potentially matter here: variation in the stimulus means, and variation in the interactions between the stimuli and the condition effect (see Westfall et al., 2014). In this article, we refer to these both simply as "stimulus variation," and we denote them both using V_S, as explained in the ***Ensuring high statistical power*** section.

References

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. doi:10.1016/j.jml.2007.12.005

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*, 575–589. doi:10.1016/S0022-5371(75)80045-4

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. doi:10.1177/1745691612459060

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. doi:10.1016/j.jesp.2013.10.005

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376. doi:10.1038/nrn3475

Cartwright, D. (1971). Risk taking by individuals and groups: An assessment of research employing choice dilemmas. *Journal of Personality and Social Psychology*, *20*, 361–378. doi:10.1037/h0031912

Cartwright, D. (1973). Determinants of scientific progress: The case of research on the risky shift. *American Psychologist*, *28*, 222–231. doi:10.1037/h0034445

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359. doi:10.1016/S0022-5371(73)80014-3

Coleman, B. E. (1964). Generalizing to a language population. *Psychological Reports*, *14*, 219–226. doi:10.2466/pr0.1964.14.1.219

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

Enserink, M. (2012). Final report on Stapel also blames field as a whole. *Science*, *338*, 1270–1271. doi:10.1126/science.338.6112.1270

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

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*, 975–991. doi:10.3758/s13423-012-0322-y

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*, 3–12. doi:10.1177/1088868313507536

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, *61*, 23–29. doi:10.1037/h0040561

IJzerman, H., Brandt, M. J., & Van Wolferen, J. (2013). Rejoice! In replication. *European Journal of Personality*, *27*, 128–129. doi:10.1016/j.jesp.2013.10.005

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648. doi:10.1097/EDE.0b013e31818131e7

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

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

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*, 54–69. doi:10.1037/a0028347

Kenny, D. A. (1985). Quantitative methods for social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd ed., Vol. 1, pp. 487–508). New York, NY: Random House.

Kogan, N., & Wallach, M. A. (1964). *Risk taking: A study in cognition and personality*. Oxford, England: Holt, Rinehart & Winston.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*, 608–614. doi:10.1177/1745691612462586

Monin, B., & Oppenheimer, D. M. (2014). The limits of direct replications and the virtues of stimulus sampling. *Social Psychology*, *45*, 299–300.

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, *83*, 602–627. doi:10.1037/0033-2909.83.4.602

Nosek, B. A., & Lakens, D. (Eds.). (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. doi:10.1177/1745691612459058

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660. doi:10.1177/1745691612462588

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536. doi:10.1177/1745691612463401

Pruitt, D. G. (1971). Choice shifts in group discussion: An introductory review. *Journal of Personality and Social Psychology*, *20*, 339–360. doi:10.1037/h0031922

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551–566. doi:10.1037/a0029487

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80. doi:10.1177/1745691613514755

Simonsohn, U. (2013). *Small telescopes: Detectability and the evaluation of replication results* (SSRN Scholarly Paper No. ID 2259879). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2259879

Stoner, J. A. F. (1961). *A comparison of individual and group decisions involving risk* (Master's thesis). Massachusetts

Institute of Technology. Retrieved from http://dspace.mit.edu/handle/1721.1/11330

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

Verhagen, A. J., & Wagenmakers, E. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475. doi:10.1037/a0036731

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*, 1115–1125. doi:10.1177/01461672992512005

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*, 2020–2045.