This version: 2024 11 13
Latest: https://urisohn.com/44

Stimulus Sampling Reimagined: Designing Experiments with Mix-and-Match, Analyzing Results with Stimulus Plots

Uri Simonsohn
ESADE Business School
urisohn@gmail.com

Andres Montealegre
Cornell University
am2849@cornell.edu

Ioannis Evangelidis
ESADE Business School
ioannis.evangelidis@esade.edu

ABSTRACT.
Psychology experimenters choose stimuli to indirectly manipulate conceptual variables (e.g., trust, impatience, and arousal). Stimulus selection is typically unsystematic, undocumented, and irreproducible. This makes confounds likely to arise. Study results, in turn, are typically reported at the aggregate level, averaging across stimuli. This makes confounds unlikely to be detected. Here we propose changing both the design and analysis of psychology experiments. We introduce "Mix-and-Match", a procedure to systematically and reproducibly stratify-sample stimuli, and "Stimulus Plots", a visualization to report stimulus-level results, contrasting observed with expected variation. We apply both innovations to published studies demonstrating how things would be different with our reimagined approach to stimulus sampling.

Data and code to reproduce all results are available from:
https://researchbox.org/2257 (use code CXUWHS)

It is tempting to assume that random assignment justifies making causal claims based on experimental results. This, however, is generally not the case, at least not for the causal claims of interest to researchers. The reason is that randomly assigned conditions seldom differ only in the dimension of interest. For example, if you make an object heavier, you must either make it bigger or denser, you cannot *only* modify weight. Thus, if in an experiment we randomly modified the weight of an object, any observed impact, while causally attributable to the performed modification, cannot be unambiguously attributed to the change in *weight*.

This general challenge to causal inference from experimental results is particularly relevant to psychology, where many experiments attempt to manipulate conceptual variables (e.g., sadness, loneliness) by assigning participants to different stimuli seeking to *indirectly* influence them (e.g., watching a sad video, or playing a rigged game with ostracizing partners). The indirect nature of the manipulation opens up experiments to a large and difficult to exhaustively examine set of potential confounds.

For example, in his influential article on the analysis of experiments with multiple (word) stimuli, Clark (1973) discusses experiments by Rubenstein, Lewis, and Rubenstein (1971) which contrasted how long it took participants to recognize words as valid, when the words had homophones (e.g., 'maid' , 'made') vs when they did not (e.g., 'pest'). Clark noted that words have many attributes that impact how long it takes to recognize them as valid, such as length, meaning, spelling difficulty, etc. Comparisons between words with and without homophones are confounded.

Rubenstein et al. randomly assigned participants to words with vs without homophones, but obviously did not randomly assign words to have or not have a homophone, thus the correlation between whether a word has a homophone and participants' time to recognize it is just that, a correlation; one which does not warrant causal interpretation, because words with and without homophones likely differ on other dimensions too.

Clark (1973) proposed, as have many methodologists in the decades since (e.g., Baribault et al., 2018; Judd, Westfall, & Kenny, 2012, 2017; Wells & Windschitl, 1999), that the way around this problem involves using many rather than few stimuli.[1] The idea is that selecting a large enough sample of stimuli will guard against the possibility that the results are due to the particular stimuli that were chosen. This recommendation follows from these authors having diagnosed the issue as a problem of external validity.[2]

We propose here that external validity is the wrong diagnosis.

We believe the issue is not whether the stimuli that were chosen have the same effect as do the stimuli that were not chosen, but rather, whether the stimuli that were chosen have an effect *for the hypothesized reason*. The correct diagnosis, in our view, is that poorly selected stimuli, whether few or many, challenge internal rather than external validity.

Once we accept that diagnosis, that the challenge is to internal validity, the approach to choosing stimuli, to analyzing data from experiments with multiple stimuli, and to interpreting those results, changes. So, *everything* changes.

[1] This literature, in turn, is related to an earlier debate in psychology on whether it is important for paradigms and stimuli to be ecologically valid by representing the context in which the studied phenomena occur. See for instance the article by Brunswik (1955) and the rest of the special issue published in *Psychological Review* V62(3).

[2] Wells and Windschitl (1999) write "Commonly, stimulus sampling is treated as an issue of external validity in which the question is whether the results can be generalized across other participants, stimuli, times, settings, and so on. Here, we emphasize how failure to sample stimuli can threaten construct validity." (p.1116), they define construct validity quoting (Campbell & Cook, 1979), as being threatened when "the operations which are meant to represent a cause or effect can be construed in terms of more than one construct". We see this definition as ambiguous, for it is unclear whether the concern is related to whether the single stimulus used in an experiment may show an effect for a reason other than the hypothesized one, vs whether its effect generalizes to other stimuli that could have been chosen. In Supplement 5, we document that all arguments in the article by Wells & Windschitl (except for their footnote 9) involve the latter interpretation, which we see as effectively indistinguishable from external validity. The concern Wells & Windschitl had was that studies with a single stimulus may choose an *atypical* one (it's not clear to us that samples of n=5 should be expected to be more *typical* than samples of n=1; it depends on how they are sampled).

Let's focus first on that consensual view we challenge here, the need to run *many* stimuli (influential papers have proposed 20, 50, or even 100s of them).[3] The *number* of stimuli used in an experiment *does not* actually matter very much for internal validity. There is no reason to expect that, in the population of all words, those with vs without homophones are matched on all confounds that impact how easy it is to recognize a word (e.g., that they have the same average length, the same average pronounceability, etc.). Therefore, there is no reason to expect that a sufficiently large sample of words with vs without a homophone differ, even on average, only in having a homophone. There is no reason for the first 10 words Rubenstein et al. chose to be more biased than the next 10 words, nor to expect the bias of the first 10 words to cancel out the bias of the next 10. A sample of 10 basketball players over-estimates human height. A sample of 1000 basketball players does also.

Even if Rubenstein et al. (1971) had included every word in the English Oxford Dictionary as stimuli in their study, the causal inference problem would remain *unchanged*. We still would not know if observed differences between all words with vs. all words without a homophone occur *because* some words have homophones. To address bias, we don't need much bigger samples of stimuli, we need much better samples of stimuli.

It's useful to consider why a single stimulus per condition isn't usually enough to provide internally valid results. In theory, if we were certain that the only difference within a single pair of stimuli in an experiment was the intended one, then one stimulus per condition would be enough for internal validity purposes. In practice, however, we can almost never be confident of that. Running more stimuli, say 3, 5 or 10 of them per condition, can alert us to the presence of unexpected confounds, by exposing unexpected variation in effects across stimuli. When the focus

[3] Clark (1973) calls for many more than 20 words as stimuli, Judd et al. (2012) for 30 or 50 or more stimuli, Baribault et al. (2018) considers experiments with 100s of stimuli.

is on internal validity, then, we do not run more stimuli to obtain a more diagnostic mean, we run more stimuli to obtain diagnostic variation. Diagnostic of unexpected confounds.

While samples of stimuli can be too small for internal validity purposes, they can also be too large. One reason is that as the sample of stimuli grows, so does the amount of random variation among them in any given sample, obfuscating potential true differences in effects. This may seem counterintuitive, but it is simply a multiple-comparisons problem. The more stimuli a study has, the more likely some will differ from others by chance, and thus the harder it becomes to diagnose a given observed difference across stimuli as *not* arising from chance.[4] Another reason why large samples of stimuli may lower internal validity is that generating stimuli that are free of confounds is difficult, and generating many stimuli that are free of confounds is necessarily more difficult.

Thus, once a study has 5 or 10 stimuli per condition that are meaningfully diverse there may be a limited benefit of additional stimuli from an internal validity perspective. We are not advocating against large sets of stimuli, rather, we are pointing out that for *internal validity* purposes large sets of stimuli are neither necessary nor sufficient.

A key realization is that stimuli are typically the means, not the end. Rubenstein et al. cared about how language is encoded and retrieved by people, they did not care about the average time it takes to recognize a homophone as a valid word; probably nobody cares about that. Because psychology experiments rely on stimuli to operationalize conceptual variables, the stimuli are not usually of intrinsic value (though they can be in some settings, e.g., we may intrinsically value how people evaluate a specific piece of fake news, or a specific government policy).

[4] Note that if there is a specific hypothesis for how stimuli will differ, for example if a stimulus attribute acts as a moderator, then a larger number of stimuli may be necessary to test the hypothesis. Since the heterogeneity being examined would no longer be fully exploratory, a larger number of stimuli (with different moderator values) would not be harmful for the purposes of checking internal validity.

We now switch our working example from homophones to disgusting videos. Several experimenters have examined the causal impact of incidental disgust by having participants watch a toilet scene from the film "Trainspotting", sometimes using sadness as a control condition, e.g., watching a scene from the film "The Champ", where a kid cries over his dead father's body.[5] If these two scenes differed on anything other than the disgusting aspects of the Trainspotting scene, which they obviously do, the disgust manipulation would be confounded. Again, we randomly assign participants to watch a clip, we don't randomly assign a given movie scene to be disgusting. And, again, simply collecting a large sample of stimuli does not solve the problem, for there is no reason to expect that, on average, disgusting and non-disgusting scenes are matched on all (or any) other attribute that could impact moral judgments. Figure 1 depicts this situation, showing two of many possible confounds in each condition. And again, psychologists do not run studies with disgusting scenes to estimate the average effect of all possible disgusting scenes they could have chosen. Instead, they run studies with disgusting scenes to assess how the mind reacts to experiencing disgust through an (assumed to be) clean manipulation of disgust.

[5] Landy and Goodwin (2015), identify four articles that have used the Trainspotting clip to induce disgust in the context of moral judgments. In addition, Lerner, Small, and Loewenstein (2004) use it in an endowment effect study.

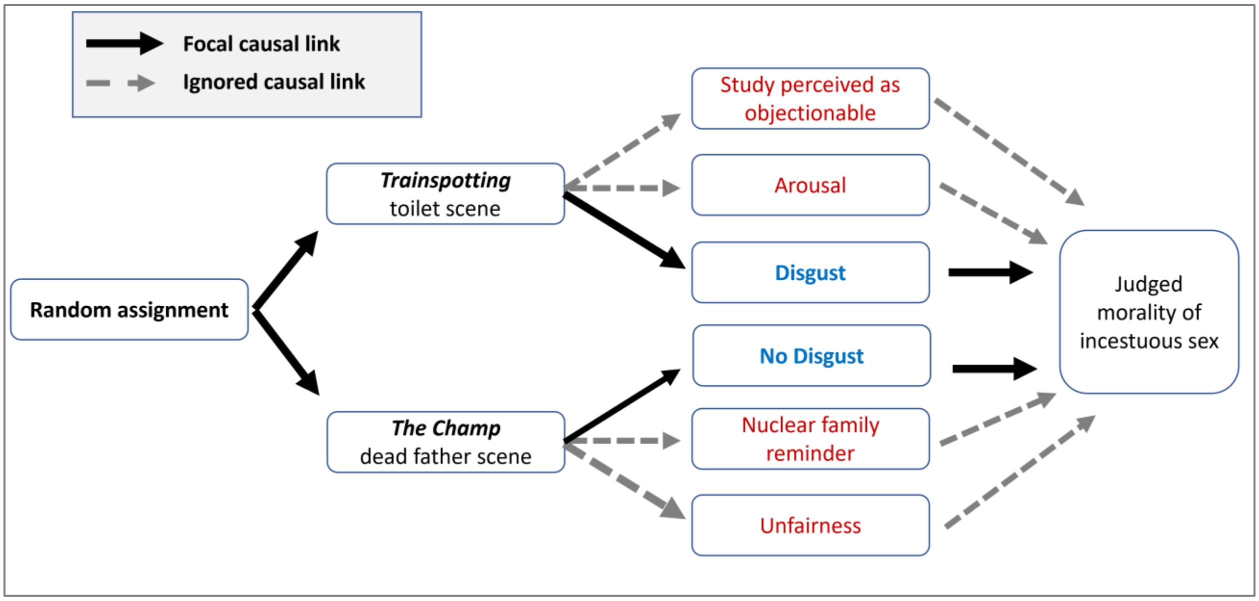

Focal causal link
Ignored causal link
Random assignment
Trainspotting toilet scene
The Champ dead father scene
Study perceived as objectionable
Arousal
Disgust
No Disgust
Nuclear family reminder
Unfairness
Judged morality of incestuous sex

Figure 1. Example of focal vs confounded causal links in psychology experiments

In light of this fundamental and ubiquitous challenge to the validity of psychology experiments posed by the fact that stimuli are often confounded, we believe confound management should be at the center of experimental design and analysis.

Stimulus Sampling Reimagined

In this paper, we reimagine stimulus sampling, the selection of stimuli for a given study (Wells & Windschitl, 1999), focusing on confound management. We propose (1) a concrete procedure for choosing stimuli and (2) a simple approach for analyzing stimulus-level results. We believe both are applicable to most psychology experiments.

In terms of study design: reading papers today, one seldom knows why the specific stimuli used were selected, how they were selected, and what other stimuli the authors would have considered valid (or invalid) substitutes. Papers often discuss confounds of chosen stimuli as afterthoughts that motivate the next study, or in the Limitations sections, or perhaps more often, not at all. Our proposal for generating stimuli, Mix-and-Match, changes all of this.

Mix-and-Match is a systematic, documentable, and reproducible process of stimuli generation which helps researchers be transparent about how and why they operationalize their conceptual variables with the chosen stimuli, disclosing the confounds they considered, and how they attempted to address them. Confound management is moved to the earliest part of the discussion of experiments: the design section.

In terms of study results: reading papers today with multiple stimuli, one seldom learns about effects at the individual stimulus level. Results, instead, are reported at the aggregate level, often relying on mixed-models which control for, but do not expose, variation across stimuli (see e.g., McNeish, 2023).[6] Our proposal of constructing "Stimulus Plots" changes all of this.

Stimulus Plots depict results at the individual stimulus level, helping authors and readers identify which stimuli do and do not show the effect, and which contribute more or less than expected to the overall average. We demonstrate the use and contribution of Stimulus Plots re-analyzing data from recently published papers showing examples when the conclusions do, and do not come into question when variation across stimuli is considered.

We write this paper with four main goals: (1) that researchers who run studies with only one stimulus per condition, will consider running them with a few stimuli instead, (2) that researchers who run studies with multiple stimuli, will more purposefully, systematically and transparently choose their stimuli (using Mix-and-Match), (3) that authors and readers will no longer act as if internal (or external) validity have been addressed by the mere fact that a significant overall result is obtained having used many stimuli, and (4) that authors and readers of studies with multiple stimuli will actively explore variation in the results across carefully chosen stimuli, through Stimulus Plots, to explicitly assess internal validity. We believe our proposals apply to all

[6] The output of a mixed model *can* be used to explore variation. In R, after a mixed-model is estimated with model1=lme4::lmer(), the random effect estimates can revealed with *ranef(model)*.

behavioral experiments where internal validity is important, that is, where establishing why an observed effect is attained is important.

The remainder of the article is organized as follows. First, we distinguish among three types of experimental designs based on how stimuli are selected. Next, we introduce Stimulus Plots, illustrating their use by reanalyzing data from three published papers. Following this, we present Mix-and-Match, our proposed procedure for selecting stimuli. We then preview what we envision the future of stimulus sampling in psychology to be. In the general discussion, we answer a series of questions we imagine some readers may have, such as, "isn't external validity also important?", "doesn't mediation take care of internal validity?" and "does using multiple stimuli reduce statistical power?" We close by discussing limitations we see in our proposals. The online supplement covers various issues in detail that may be important for some but not other readers.

Three experimental designs

Throughout this article we distinguish among three types of experimental designs: (i) treated-stimulus, (ii) matched-stimulus, and (iii) compared-stimulus designs. In treated-stimulus designs, stimuli are selected for one condition, and they are treated (modified) to be used in the other condition (e.g., participants evaluate the same news story in one condition with a fact-check, and in the other condition without a fact check). In matched-stimulus designs, stimuli are sampled separately for each condition and are then matched forming pairs of similar stimuli across conditions (e.g., comparing reactions to pairs of real vs fake stories where each pair contain stories with similar attributes). Lastly, in compared-stimulus designs, stimuli are sampled separately for each condition, and the entire sets are compared without matching individual stimuli across conditions (e.g., comparing the average reaction to a set of true vs a set of fake stories). Achieving internal validity is easiest with treated-stimulus designs and hardest with compared-stimulus designs.[7]

Stimulus Plots

Only by analyzing data at the individual stimulus level can the main goal of stimulus sampling be achieved: assessing internal validity.[8] Estimates are necessarily noisier when based on subsets of data, therefore, the expectation should not be that every stimulus is individually statistically (or practically) significant, or even that all estimates have the same sign. Even if stimuli had the same true effect, because of sampling error, different stimuli will have different effect size estimates. Rather than conducting confirmatory analysis on each individual stimulus, the idea is to conduct exploratory analysis across them. To enable answering questions like: Is the effect evident only for a small subset of stimuli? Does a surprising share of stimuli show an effect in the opposite direction? Are there outlier stimuli with surprisingly big or small effects that may shed light on confounds?

We propose analyzing individual stimuli relying on what we refer to as "Stimulus Plots", plotting stimuli-level results side-by-side. For treated- and matched-stimulus designs, Stimulus Plots can have two panels: one plots the means for each stimulus in both conditions, the other the differences of means for each stimulus across conditions (note: proportions are also means). For compared-stimulus design only means can be depicted, as there is no stimulus-level effect. As we

[7] In treated stimulus designs it is possible for the treatment to alter how the 'same' stimulus is perceived, creating a confound. For example, adding a fact-check to a story may alter how participants interpret ambiguous claims or word-choices. For a recent example in a different domain where a treated-stimulus design produces a confound through an interaction, see Spiller (in press).

[8] We have come across some papers that report stimulus-level results (see e.g., Bar-Hillel, Maharshak, Moshinsky, & Nofech, 2012; Dias & Lelkes, 2022; Evangelidis, Levav, & Simonson, 2023; Novoa, Echelbarger, Gelman, & Gelman, 2023). But, it does not seem that this was done with the goal of assessing internal validity, and they did not contrast observed with expected variation. We believe our proposed Stimulus Plots would have added to the informativeness of even these papers that already reported stimulus-level results.

show below, when discussing Example 2, we propose a different kind of plot for compared-stimulus designs (a beeswarm plot), which reflects the unpaired nature of the stimuli across conditions.

While stimulus plots are exploratory, we propose visually contrasting the observed heterogeneity of effect size across stimuli, with that which would be expected if all stimuli had the same effect size (under 'homogeneity'). This contrast helps calibrate the meaningfulness of differences in observed effect sizes, preventing researchers from over-interpreting random noise, and assessing if a pattern of interest is actually surprising. We provide an R package, 'stimulus', with a function, *stimulus.plot(),* that allows users to create publication-ready Stimulus Plots running a single line of code. We next illustrate the use of Stimulus Plots by re-analyzing data from three recent papers.

Example 1. Some stimuli show no effect, some show huge effects

In their Study 4, Salerno and Slepian (2022) examine whether people report that revealing another person's secret as punishment is more acceptable when the secret involves an intentional rather than unintentional transgression. The authors created 20 vignette pairs. Each pair involved an intentional and an unintentional version of a similar act. For example, in one vignette (referred to as *'drug'* in our Figure 2 below), the intentional version reads "*Ross brought illegal party drugs to a party, which he then took when he got there.*", while the unintentional one reads "*Ross went to a party, and although he had decided beforehand, he would not take any illegal party drugs, a friend offered him some, and in the heat of the moment, he said yes.*" (see their Appendix C; p.24). This design is in between a treated-stimulus design and a matched-stimulus design, in that a given story has two versions (the story was treated), but the treatment (of intentionality) is often quite rich, modifying the underlying context often well beyond intentionality. For some stimuli pairs, we can more accurately think of them as two stories that were arguably matched rather than a single story that was treated.

The authors report, as is customary, only the overall effect across all 20 stimuli pairs: higher average acceptability of revealing secrets of intentional acts, M_1=2.55 vs M_2=3.20, p<.001, which we reproduced using their posted data. However, we also ran an ANOVA heterogeneity test (see e.g., McNeish, 2023) which suggests significant variation in effect size across stimuli, $\chi^2(2)=74.24$, $p < .0001$.[9] We can explore this variation with our proposed Stimulus Plots, depicted in Figure 2 (to be clear, we advocate reporting Stimulus Plots even in the absence of statistically significant heterogeneity).

The left panel shows the means for each stimulus-pair separately by condition, while the right panel displays the corresponding mean differences across conditions, along with the confidence intervals obtained from simple t-tests run on each stimulus. These inform the precision of the estimates for each stimulus (e.g., whether they are individually statistically significant).

The right panel also depicts how much heterogeneity in effect size across stimuli we should expect from chance alone. Specifically, the overall average effect is M_1=2.54 vs. M_2=3.21, a difference of 0.67. If all 20 stimuli had a true effect of 0.67, in any given study, some effects would be estimated above 0.67 and some below. The dashed line shows how much above or below 0.67 we should typically expect different results to be (if your intuition is that the dashed line should be flat, see footnote).[10] The light-blue confidence band displays the range of deviations from

[9] The test compares a mixed-effects model with only random intercepts for stimuli, with one that has both random intercepts and random slopes; if it's significant, the random slopes model fits the data significantly better.

[10] Here we explain why the expected effect size line in Stimulus Plots is not flat. Consider a simple example where 100 people toss 10 fair coins each. We wouldn't expect all 100 to toss 5 heads and 5 tails—some will toss more heads than others. In fact, we expect the top head-tosser to get about 9 heads, and the bottom one just 1 head. The same logic applies to effects sizes for stimuli. If the true effect is 0.67, we don't expect every stimulus to obtain a 0.67 effect in any given sample, some will be above and some below 0.67. Through resampling we compute how much above and below 0.67 we should expect each ranked stimulus to be.

expectations that is expected with 95% confidence. Note that the confidence band is not about the specific stimulus per-se, but about ranked stimuli. It tells us how extreme we would expect the biggest effect to be, the second effect to be, etc. In any random sample which stimulus happens to be the highest is random. These calculations are done relying on resampling. See Supplement 9 for technical details, including the intuition for why the confidence band is narrower than are the confidence intervals.

If all observed effects fall within or near the 95% confidence band, the level of variation in effects across stimuli is consistent with pure chance (so the heterogeneity test would not be significant), while if many effects fall outside the 95% confidence band, there is more variation than expected by chance alone (and the heterogeneity test would tend to be significant). The Stimulus Plot reports a formal test of heterogeneity based on this comparison, related to, but distinct, from the ANOVA test reported earlier; see figure legend.

A key pattern in the right panel, then, is that in this study the level of variation in effects across stimuli is much larger than would be expected by chance, since most dots are well outside the 95% confidence band. Both the Stimulus Plot and the χ^2 heterogeneity test reported above, then, tell us that the effects differ by more than expected by chance, but only the Stimulus Plot allows us to learn about the nature of the heterogeneity. Here, for instance, we see that there are several stimuli showing essentially no effect, while a few stimuli show effects that are substantially bigger than expected.

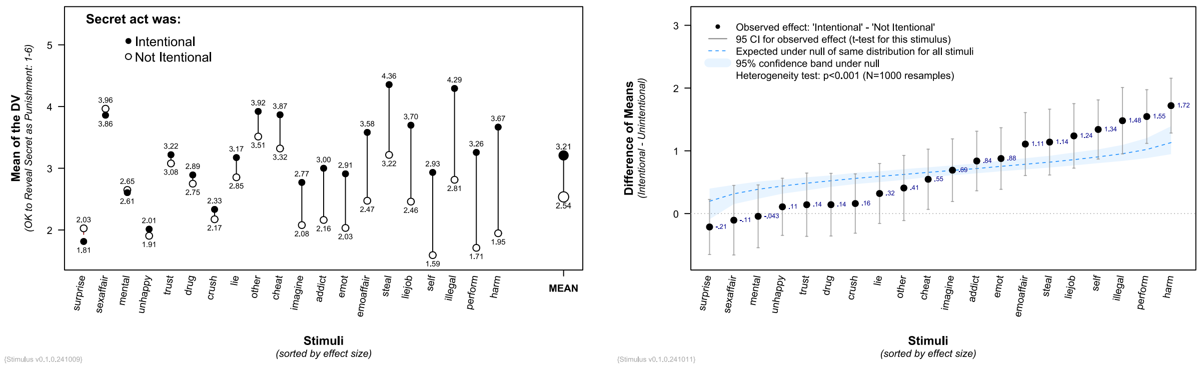


Figure 2. Stimulus Plots for Study 4 in Salerno & Slepian (2022)

The study involves a 2-cell treated-stimulus design, comparing participants' willingness to reveal another person's secret based on whether the transgression was intentional or not intentional. The expected line and its 95% confidence interval in the right panel are obtained via resampling, by recomputing the average difference of means for each stimulus after shuffling the stimulus label across rows repeatedly.

R Code to reproduce the figure: https://researchbox.org/2257/48 (code CXUWHS)

That a substantial share of stimuli "do not work" in this study does *not* necessarily invalidate its main conclusion (indeed, 11 of the 20 stimuli are individually significant), but it does warrant a deeper exploration of the design and results. For example, are there moderators or confounds that may explain why the effect is so large for some stimuli while absent from several others?

Figure 2 drew our attention to the vignette leading to the largest effect, "harm", which involves John cutting himself intentionally ("to deal with his emotional pain"), vs unintentionally ("while chopping vegetables").[11] We wondered whether the large difference in willingness to reveal that John cut himself across conditions may arise because respondents wished to *help* John with his self-cutting problems rather than to *punish* him.

Our attention was also drawn to the vignette with the smallest effect (directional reversal), "surprise", which involved Kathy surprising her husband with opera tickets intentionally ("kept this a surprise for months") or unintentionally ("had forgotten to put it on their shared calendar").

[11] See Appendix C in Salerno & Slepian 2022, p.24.

We wondered whether the directional reversal may arise because the action isn't immoral whether intentional or unintentional, and intentionality may make it a more *positive* act.

All of this is speculative of course. But speculation is the goal of Stimulus Plots. Generating hypotheses about surprising variation in effect size that can be explored with more data either before or after the work gets published.[12]

This example also illustrates that authors who are following the current consensus advice for stimulus-sampling by including a large number of stimuli and analyzing the results with mixed models are not addressing the concerns we raise here regarding internal validity.[13] We are not pointing the finger at these authors or this study, we are pointing the finger at the current consensus.

Example 2. Many stimuli show significant reversals

Karmali and Kawakami (2023) examine differences in how Black vs White people are perceived when assuming expansive vs constrictive poses (i.e., 'power posing'). Their paper reports 4 studies, all relying on the same photographs of 20 Black and 20 White men assuming two different expansive and two different constrictive poses.[14] We focus on Study 3, where n=105 undergraduates chose potential partners for an upcoming task. They each saw 20 sets of 4 photographs of different people (crossing race and pose within each set of four) and they chose one out of the four as a potential partner. The pose manipulation (target assumes a constrictive vs

[12] We analyze all 20 stimuli for potential confounds in Supplement 10.
[13] Salerno & Slepian write "By modeling the content of the secrets as a random category . . ., we can conceptually generalize the current results to the larger universe of unsampled secrets . . . (Judd et al., 2012)." (p.623).
[14] The design involves 5 expansive and 5 constrictive poses. Any given potential target was shown in 2 out of 5 poses of each kind.

expansive pose) involves a treated-stimulus design, whereas the race manipulation (target is White vs Black) involves a compared-stimulus design.[15]

The study's key finding is that White partners were chosen more often when in an expansive than constrictive pose (Z=4.96, p<.001), but that this effect of pose was not observed for Black partners (Z=1.26, p=.208); a race *x* pose attenuated interaction (Z=2.47, *p*=.013). The authors write that "expansive versus constrictive poses **did not influence** participants' willingness to interact with Black targets"(p.59, bold added). We obtained their posted data, reproduced this result, and then constructed Stimulus Plots (see Figures 3 and 4).

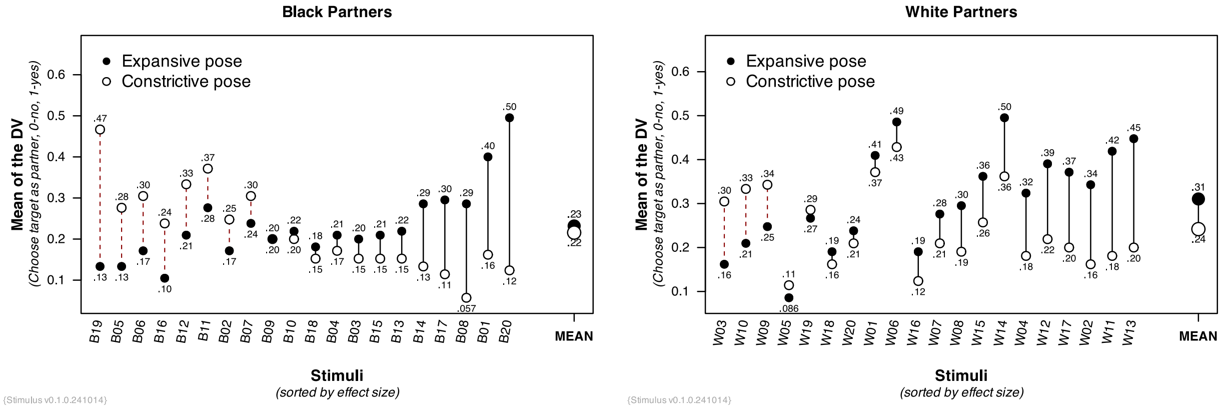


Figure 3. Stimulus Plot for Means in Karmali & Kawakami – Study 3

The study involves a 2 (race [compared]) x 2 (power posing [treated]) stimuli design. Participants chose 1 of 4 potential partners based on photographs where they were either in an expansive or a constrictive pose. The figure depicts the percentage of times each stimulus (potential partner) was chosen.

R Code to reproduce the figure: https://researchbox.org/2257/49 (use code CXUWHS)

[15] Before the study, the authors ensured that the White and Black targets were roughly similar on perceived age, attractiveness, and objective size (p.53). However, rather than forming target pairs that were matched on these attributes, they checked whether, *on average*, all Black vs all White targets were similar on these attributes. As a result, the race manipulation follows a compared- rather than a matched-stimulus design.

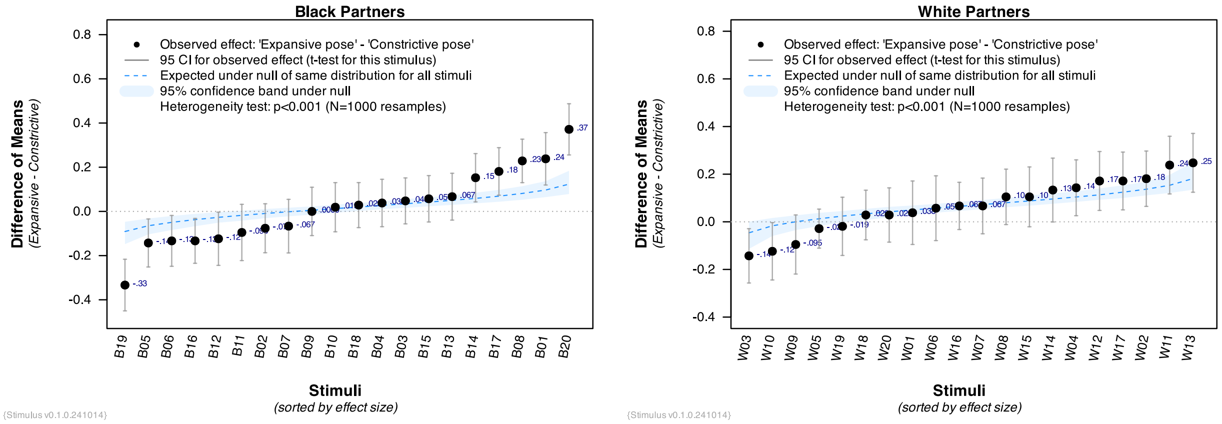


Figure 4. Stimulus Plot for Effects in Karmali & Kawakami – Study 3

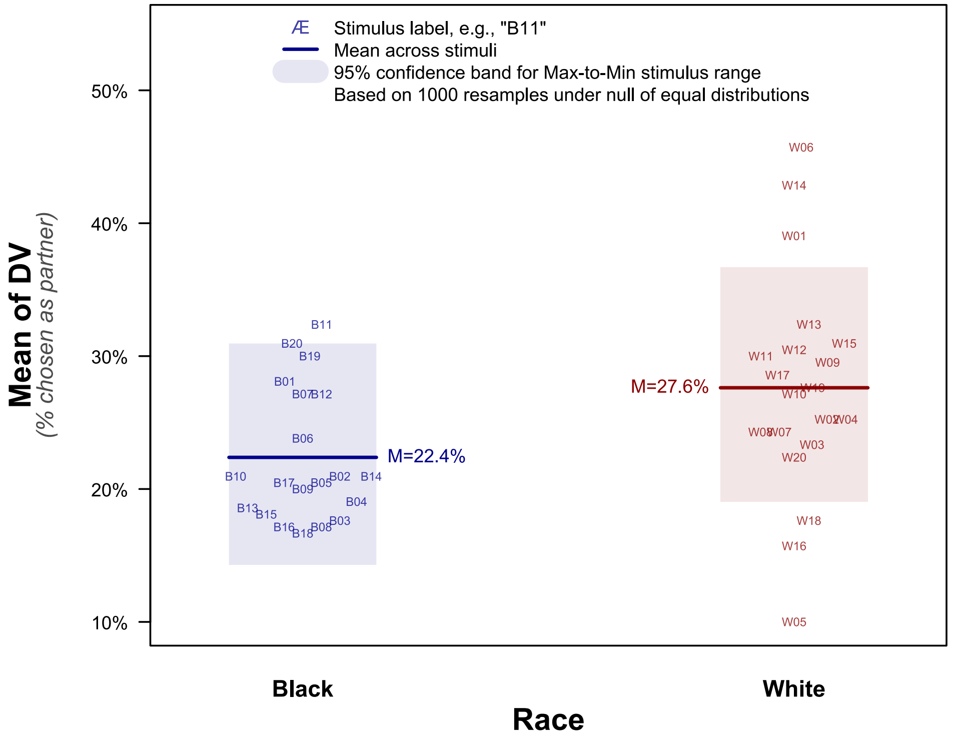
Differences computed off means from Figure 5. The expected line, and its 95% confidence interval, are obtained via resampling. R Code to reproduce the figure: https://researchbox.org/2257/49 (use code CXUWHS)

Figures 3 and 4 show that the effects are highly heterogeneous across stimuli within race, with some Black targets exhibiting significant *negative* differences across poses, others exhibiting significant positive differences, and these opposing effects cancel out on average. Indeed, in absolute terms, Black targets show directionally *larger* effects (12.7 vs 11.0 percentage point difference). This directly contradicts the conclusion in the paper that "expansive versus constrictive poses **did not influence** participants' willingness to interact with Black targets".

The key takeaway of this second example is that once heterogeneity in effects across stimuli are taken into account, the key conclusion of a study can be shown to be contradicted by the data. Moreover, the average effect for Black partners is uninterpretable until an explanation is found for why the effect is positive for some and negative for others.

As mentioned, this study had a treated-stimulus aspect, pose, and a compared-stimulus one, race. For compared-stimulus contrasts we propose relying on *beeswarm* plots; see Figure 5. While this contrast was not of interest to the authors of the article, we use their data to showcase how compared-stimulus designs can be visually analyzed. Beeswarm plots show stimuli individually rather than in pairs. They facilitate spotting heterogeneity overall, and individual stimuli behind such overall heterogeneity. The Stimulus Beeswarm Plot also includes a confidence band depicting the range of expected variation across all stimuli, if the true effect were homogeneous.

That band is also obtained via resampling. If all stimuli had the same true means, then only 5% of studies would include 1 or more stimuli outside it. Here we see multiple stimuli well outside it, implying sizeable heterogeneity. Had the authors posted the stimuli, we would want to explore possible explanations for the surprising popularity of targets W01, W14 and W06, and unpopularity of W05, W16 and W18.



{Stimulus v0.1.0.241023}

Figure 5. Stimulus *Beeswarm* Plot for Karmali & Kawakami – Study 3
The figure depicts the proportion of times each potential partner/stimulus was chosen from a set of four (overall mean is 25%). Each label depicts means for a single stimulus (e.g., W06 is the individual who was most often chosen as a potential partner, M=46%). The means aggregate for each potential partner across the expansive and constrictive poses, the figure focuses on the compared-stimulus design of the study. The colored regions are 95% confidence bands for the range of values expected between the highest and lowest stimulus, under homogeneity.
R Code to reproduce the figure: https://researchbox.org/2257/49 (use code CXUWHS)

Example 3. All stimuli seem consistent

Pretus et al. (2023) examine the psychological processes that underlie misinformation sharing. In Experiment 2 they asked N=797 participants how likely they would be to share a tweet, (which contained misinformation) on a 1-6 Likert scale. The authors relied on 16 different tweets, and the manipulation of interest to us is whether the tweet was accompanied by a Twitter fact-check message (their design is more complex and includes additional manipulated and measured differences). On this manipulation we are focusing on, the study involved a treated-stimulus design, the same story had or did not have a fact-check. The paper reports an overall average effect of the fact-check of 0.16, *p*=.006 (p.3124).

Relying on data provided by the authors upon request (they had posted the data, but not with individual stimuli identifiers), we reproduced their results and constructed the Stimulus Plots reported in Figure 6. The left panel shows some variation in effect size across stimuli, but the right panel shows that the observed variation is consistent with sampling error; also consistent with the results of the resampling-based heterogeneity test, *p*=.427.

It's worth distinguishing statistical vs practical significance here. That the observed heterogeneity is not statistically significant does not mean that it is not (potentially) substantively significant. If upon plotting a Stimulus Plot the differences in effects across stimuli were large from a practical/theoretical perspective, then what the non-significant result would tell us is not that there is no heterogeneity, but rather, that to study heterogeneity for these stimuli one needs a larger sample of participants.

That is indeed our interpretation of the results for this study, they are inconclusive, as we cannot rule out sizeable effects in either direction. The confidence band does not rule out *positive effects* up to four times larger than the observed average effect of 0.15 (see top-right), nor *negative* *effects* up to three times larger in magnitude (see bottom-left). We have absence of evidence of heterogeneity rather than evidence of its absence.

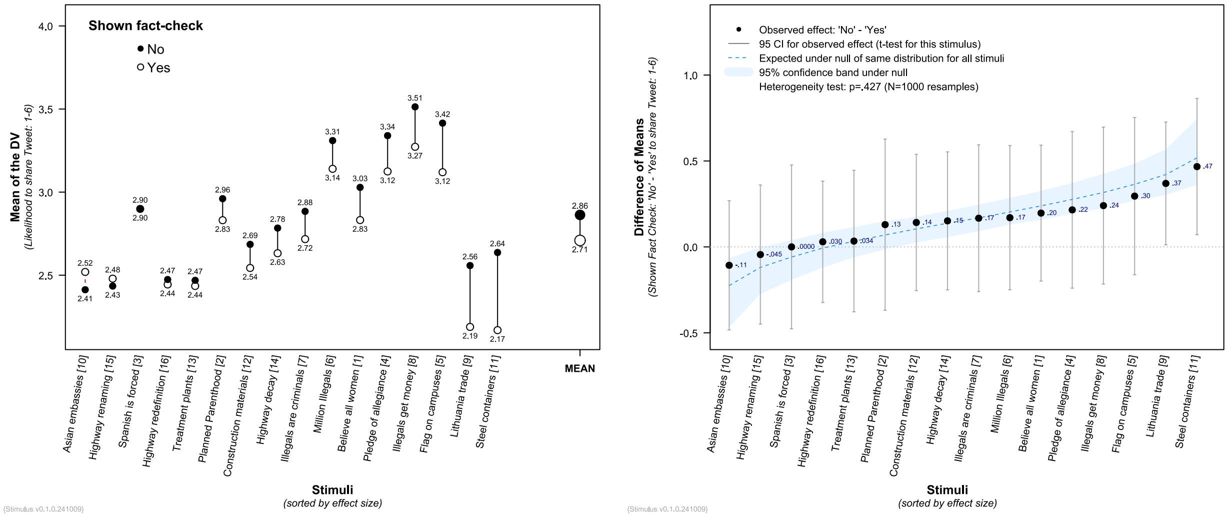


Figure 6. Stimulus Plots for Pretus et al. (2023) – Study 2

The study involves a two-cell treated-stimulus design, comparing participants' reported willingness to share a tweet containing information having been presented, or not, with a Twitter fact-check. The expected line, and its 95% confidence interval, are obtained via resampling.

R Code to reproduce the figure: https://researchbox.org/2257/9 (use code CXUWHS)

This last example showcases two important points. First, not all studies will exhibit significant or substantive heterogeneity across stimuli. Second, even in the absence of statistically significant heterogeneity, Stimulus Plots are useful (to differentiate evidence of absence of consequential heterogeneity across stimuli, from absence of evidence of it).

Interpreting heterogeneity: confounds & moderators

Stimulus Plots are useful for describing the nature of heterogeneity across stimuli (e.g., Are there reversals? Are some stimuli outliers? Are only half the stimuli showing an effect?). But human judgment is necessary for interpreting such patterns. Our motivating concern here is the potential presence of confounds, of stimuli producing effects for reasons other than hypothesized. But that is not the only possible explanation for heterogeneous effects. Moderation is undoubtedly a common source of heterogeneity in effect size, as it is often not even possible to administer a homogeneous 'treatment' across stimuli that are substantively different (e.g., as pointed out by a reviewer of this paper, the treatment of 'intentionality' is necessarily incommensurate when applied to a sexual affair vs. not studying for an exam, there is no reason to expect a homogenous effect in such a study). There may even be moderators that impact only one condition, for example, as pointed out by another member of the review team, it is possible that "a target's physical size changes the effect of the pose for Black targets, but not White targets". Heterogeneity, then, is not intrinsically problematic.

We think researchers should pay disproportionate attention to potential confounds over potential moderators, as we do in this paper, because the conservative nature of the scientific process demands that alternative explanations be ruled out before even tentatively accepting a hypothesized explanation for a finding. Confounds must be addressed first; only then should moderators be considered. One way to frame the primacy of confounds over moderators, which may resonate with readers, is that it seems reasonable to leave it for future research to establish whether a valid finding has interesting moderators, but it does not seem reasonable to leave it for future research to establish whether an interesting finding is valid.

Mix-and-Match: Systematically Generating Stimuli for Psychology Experiments

We designed Mix-and-Match following three guiding principles. The first principle is that *stimuli* should be blind to hypothesis. It is widely accepted that *participants* should be blind to hypothesis, due in part to concerns of demand effects (see e.g., Rosenthal, 2009). But the notion that *stimuli* (selection) should be blind to hypothesis is seldom if ever considered. The concern we have in mind is that when experimenters choose stimuli, they can often mentally simulate the experiment they are designing, and anticipate whether a particular stimulus is likely "to work". At the same time, it may be difficult to anticipate *why* it may work. This can lead researchers to (possibly unintentionally) be disproportionately likely to select stimuli that work for the wrong reasons (see e.g., Strickland & Suben, 2012). If, instead, experimenters chose stimuli by following a stated and reproducible rule, the stimuli become less *individually* selectable, and thus closer to, if not strictly, blind to hypothesis. Writing down a reproducible rule for selecting stimuli is thus part of Mix-and-Match.

The second principle is that stimuli should be diverse in ways that could help diagnose overlooked confounds. This involves varying stimuli on dimensions directly related to the operationalization of the conceptual variable of interest. For example, if visual stimuli are chosen to trigger disgust, variation should be along the ways in which disgust can be triggered visually (bodily fluids, pests, rot, etc.). This is the 'mixing' in Mix-and-Match.

The third principle is that there should be an explicit and defensible reason to expect stimuli across conditions to differ only, or at least primarily, on the attribute of interest. This is the 'matching' in Mix-and-Match. From a confound management perspective, matching seeks to deal with confounds researchers anticipate, by controlling for them, and mixing seeks to deal with confounds they do not anticipate, by exploring variation across diverse stimuli.

In various stages of the Mix-and-Match process we propose ways in which generative artificial intelligence (GenAI) can be used to aid the process of generating stimuli. But Mix-and-Match does not *require* GenAI. We propose three template 'prompts' to implement Mix-and-Match, theses prompts can be given to artificial intelligence agents (like ChatGPT) but also to natural intelligence agents (like pilot participants or research assistants).

Mixing stimuli

Mixing puts the sampling in "stimulus sampling". We propose the following three-step procedure for sampling stimuli: (i) defining the "experimental paradigm" that will be used, (ii) identifying the universe(s) of stimuli that could be selected or generated for such paradigm, and (iii) stratify-sampling stimuli from those universes. Steps (i) and (ii) are performed by experimenters (who are not blind to hypothesis), while step (iii) is performed by a third party that is blind to hypothesis (which may involve a GenAI agent), following instructions written by experimenters.

(i) Identifying the experimental paradigm. We define the term 'paradigm' as the description of an experimental procedure where every design element is necessary for it to be a valid and practical test of the hypothesis of interest. For example, an experimental procedure could be described simply as "disgust will be induced, and moral judgments will be elicited." But such a level of (un)specificity allows for too diverse a set of stimuli, say, based on disgusting book passages, disgusting videos, and week-long internships in a slaughterhouse. It is *impractical* to include such a broad range of stimuli in the same experiment, thus the experimental paradigm should, for practical considerations, entail more narrowly defining how disgust will be induced. Similarly, moral judgments can be elicited over too broad a range of targets (e.g., vignettes, videos, and in-person biblical reenactments), combining such diverse set of stimuli in a single study would be impractical, thus the paradigm would specify how the ambiguously immoral behavior is presented to participants.

The experimental paradigm, then, needs to be actionably specific. Something like: "participants will read paragraph-long texts, extracted from published books, that induce either disgust or sadness, and will then evaluate the morality of an ambiguous act described in a short vignette, providing their moral judgments in a 1-(very immoral) to 7-(completely moral) scale."

We have discussed the importance of sampling stimuli for the independent variable. In terms of the dependent variable, combining results across dependent variables often imposes substantive practical challenges and thus, absent explicit interest in assessing the properties of a dependent variable, the paradigm could specify a single (rather than a set of alternative) dependent variables.

(ii) Universe of stimuli. The set(s) of stimuli that meet the description of the experimental paradigm constitutes what we refer to as the universe(s) of stimuli. In our working example, one universe of stimuli involves every passage of text, across all published books, that induces disgust on the reader. Another universe of stimuli is the infinite and uncountable set of vignettes that could be generated to describe a morally ambiguous act.

(iii) Stratify sampling. Given our emphasis on internal rather than external validity, we don't propose sampling the universe of stimuli in a representative fashion; in fact, it is often unfeasible and even meaningless to speak of representative samples from a universe with infinite, uncountable, and sometimes simply undefined units (e.g., one cannot draw a representative sample of all possible vignettes that could be written to depict a morally ambiguous act). What we propose, instead of random sampling, is stratified sampling. We propose creating possibly arbitrary strata for the universe of stimuli, strata which are meaningfully different from each other, differing on central rather than peripheral attributes. For example, if inducing disgust by text, strata can differ in the nature of the disgust being induced: sexual, rot, pest, etc., rather than in the length of the text or some other auxiliary feature. If creating strata for fake news, the strata may differ on the topics of the news (e.g., transportation, health, economics), and the nature of the fakery, rather than having only transportation stories that are fake in the same way. As a default, we propose creating 5 strata but if authors have a reason to choose a different number they should.

We do not propose a relatively small number of strata because we think this is enough to guarantee internal validity—no number of strata can do that. Instead, we suggest this because more strata require more stimuli, and studies with large numbers of stimuli have two downsides. First, from a statistical perspective, as discussed earlier, the resulting heterogeneity becomes harder to interpret, because random variation across a larger number of stimuli overwhelms possibly important differences in the signal. Second, designing clean stimuli is often difficult. Designing 20 scenario pairs that cleanly manipulate a construct of interest is quite a challenge, designing 100 or 500 of them, extremely challenging; as it would be for readers and peer-reviewers to evaluate such a stimulus-rich design. Researchers working with paradigms that circumvent these challenges for large sets of stimuli could sample more strata (e.g., in cognitive psychology studies with many observations per participant and where stimuli vary on easy to operationalize dimensions).

Continuing now with stimulus selection. From each stratum, experimenters generate (sample) a number of stimuli, we propose 1 or 2 stimuli per stratum as a default, but if authors have a reason to choose another number they should.

It is *not* a problem if the strata are not exhaustive (e.g., that the strata do not encompass all possible operationalizations of the construct), nor if different researchers would produce a different stratification. The goal, remember, is not to produce a representative sample of stimuli, the goal is to produce meaningfully diverse stimuli selected (largely) blind to hypothesis.

We next propose concrete steps to implement mixing, for categorical stimuli (e.g., scenarios) and then for numerical stimuli (e.g., probabilities and monetary amounts).

Categorical stimuli. There are multiple approaches that could be relied upon to stratify categorical stimuli, such as relying on a third party (e.g., consumer good categories at Amazon.com) or prior research (e.g., the disgust categorization by Haidt, McCauley, and Rozin

(1994)). Researchers could also conduct a pilot study, or ask research assistant blind to the hypothesis, to stratify sample the universe of interest. Another alternative is generative artificial intelligence (GenAI). We provide more details about the implementation with GenAI because it is more novel an approach.

To stratify the universe of stimuli, we propose the following stimulus sampling prompt: "*please generate 5 categories of <stimulus universe> that differ in <dimension used to create strata> and provide two specific examples of <stimuli> for each.*" (note that in the prompt we use the word 'categories' rather than 'strata'). That second placeholder, *'dimension used to create categories'*, involves specifying which aspect the stimuli should vary on, with the goal of generating stimuli that are meaningfully diverse, entailing different instantiations of the conceptual variable of interest.

Applying it to our working examples:

Prompt 1: *Please generate 5 categories of homophones that differ in their etymological origin, and provide two examples of specific homophones for each.*

Prompt 2: *Please generate 5 categories of book scenes that may induce disgust that differ in the origin of the disgust being induced, and provide two examples of specific books of fiction containing such scenes (e.g., the category of bodily fluids could contain a passage from the toilet scene in Trainspotting).*

It can help to include in the prompt an example of one stratum and stimulus, as we did in prompt 2 above.

The categories produced by ChatGPT in response to Prompt 1 involved homophones that: originate in a different language, have different roots in the same language, involve different parts of speech, have different derivational processes, and were impacted by different sound changes. The examples included: "flour/flower", "knight/night", "rays/raise", "maid/made ", and "son/sun".

Prompt 2 led to stratifying disgust by its source: bodily fluids, filth, putrefaction, gross-out horror, and moral repugnance. The examples included segments from 10 books including the following five: "The Road", "The Sisters Brothers", "The Shining", "Haunter", and "Lolita" (one for each stratum).

The stratification generated with the stimulus sampling prompt, whether provided by a natural or artificial intelligence agent, includes a random component, thus the same prompt may lead to different results over time. Moreover, different researchers may operationalize it differently. This idiosyncratic variability is again fine, as long as the stratification produces meaningfully diverse stimuli, because the goal is internal validity, not generalizability.

Numerical stimuli. The stimulus sampling prompt is useful for categorial stimuli. For numerical stimuli, for example monetary outcomes or probabilities, we propose instead, including in the paradigm definition the set of numbers that would be considered a practical and valid test of the hypothesis of interest (e.g., that to facilitate mental calculations the numerical stimuli need to be multiples of 100, but smaller than 10,000, and the probabilities should be multiples of 10% and smaller than 100%). For stratified sampling one could then choose a diverse set of numbers spanning the range of the consideration set. We exemplify this in Supplement 4, by providing a Mix-and-Match based design of the classic "Asian Disease" problem (Tversky & Kahneman, 1981).

Matching stimuli

The "Match" in Mix-and-Match involves striving to generate stimuli that across conditions differ only, or at least primarily, on the focal attribute of interest to the experimenter; striving to match stimuli on all identified potential confounds across conditions. Ideally stimuli are individually matched forming pairs, so that every stimulus in one condition is paired with a matched stimulus in another condition, providing multiple mini-replications within a study. In treated- and matched-stimulus designs stimuli are paired, whereas they are not in compared-stimulus designs.

In treated-stimulus designs, stimuli are selected for one condition, and those stimuli are either treated (modified) to be used in the other condition, or used in both conditions in the presence vs. absence of the treatment of interest. The question of whether pairs of treated/untreated stimuli differ only on the dimension of interest should be explicitly argued for by experimenters, and evaluated by readers. The "confound confirmation prompt" we propose later can be used for such purposes.

In matched-stimulus and compared-stimulus designs, stimuli are sampled separately across conditions. Examples include experiments examining how participants respond to male vs female names, experiential vs material purchases, disgusting vs sad videos, words with vs without homophones, and verbal vs math problems. These designs are naturally more challenging from an internal validity perspective than are treated-stimulus designs, because stimuli can differ on many, possibly infinite, non-focal attributes across conditions.

To match stimuli in such designs requires identifying confounding variables (ways in which the stimuli may differ in their impact on the dependent variable other than the focal mechanism), and then measuring those confounding variables for candidate stimuli. For example, for homophones, one identifies other word attributes that may influence how quickly people can recognize them as valid words, and measures those attributes: say, word frequency, language origin of the word, spelling difficulty, etc.

To identify potential confounds researchers could rely on the following 'confound exploration prompt': *"what variables might be expected to predict variation in <dependent variable> across <class of stimuli>?"*, any variable that is identified and is not the focal variable being manipulated constitutes a potential confound. For instance, for the homophones study one could ask "*what variables might be expected to predict variation in reaction time to recognize a word as valid, across different words?".* Any variable other than "being a homophone" constitutes a potential confound that should be measured and attended to.

This confound exploration prompt *can* be answered by the researchers themselves, but because they are not blind to hypothesis and they have a stake in the hypothesis, they may fail to detect consequential confounds. We thus recommend posing those questions to others who are blind to hypothesis, be it research assistants, participants in a pilot study, or a GenAI agent.

When we posed that confound exploration prompt for homophones to ChatGPT it identified 10 variables, including word length, frequency, phonological regularity, and semantic transparency. Researchers naturally need to apply expert judgment to filter the suggestions produced with this prompt. Proposed confounds may actually be mediators or simply irrelevant.

Having identified potential confounds, researchers can then measure the candidate stimuli on those attributes (e.g., with a pilot study where participants rate the stimuli). For a matched-stimulus design, pairs of stimuli across conditions are formed by matching a target stimulus, say the word "bear", to the word without a homophone that is most similar to "bear" on all measured attributes (a 'nearest neighbor' approach). If a particular target stimulus lacks a sufficiently similar control based on the measured covariates, then it probably should not be used at all; otherwise, it introduces an unsolvable confound.

Sometimes such paired-stimuli designs may be unfeasible, e.g., stimuli are not selectable or modifiable at a sufficiently granular level to allow forming pairs that differ only in the focal attribute (e.g., it may be unfeasible to create pairs of videos that differ only on whether they are sad vs disgusting). In such cases, we would recommend that experimenters consider changing the paradigm (e.g., inducing emotion with vignettes instead of videos). If the paradigm must be used (e.g., because the manipulations are of intrinsic interest, such as assessing the impact of violent videos), then we have a 'compared-stimulus' design, where a set of stimuli in one condition is compared to a set of stimuli in the other. Here experimenters may rely on a statistical model (e.g., linear regression) to control for the confounding variables. For example, this could involve running an emotion induction task using various disgust and sadness videos (say, 10 of each), and reporting the effect of disgust vs sadness controlling vs not-controlling for other attributes identified as potential confounds, measured for each video. Intuitively, one looks for *absence* of mediation for the confounds, or at least that a substantial portion of the effect survives controlling for them. We recommend that researchers use this approach only when other alternatives are unfeasible, while remaining mindful of its risks, such as measurement error making the use of controls in regression an imperfect approach to dealing with confounds (see Westfall & Yarkoni, 2016).

For treated- and matched-stimulus designs, we propose a final check to validate pairs, posing the following "Confound confirmation prompt": *We are going to describe two <stimuli>, please identify 5 consequential differences between them that may impact <the dependent variable> in <the hypothesized direction>*. If none of the 5 consequential differences are deemed plausible confounds by the experimenter, the stimuli-pair is ready for use. In some paradigms this final check may be redundant and thus unnecessary.

For example, we took the aforementioned "self-cutting" example from Study 4 in Salerno and Slepian (2022), the one exhibiting the largest effect of all, and posed the "stimulus-pair verification question" to ChatGPT through the following prompt: *"We are going to describe two scenarios, please identify 5 consequential differences between them that may lead people to be more prone to sharing scenario 2.*

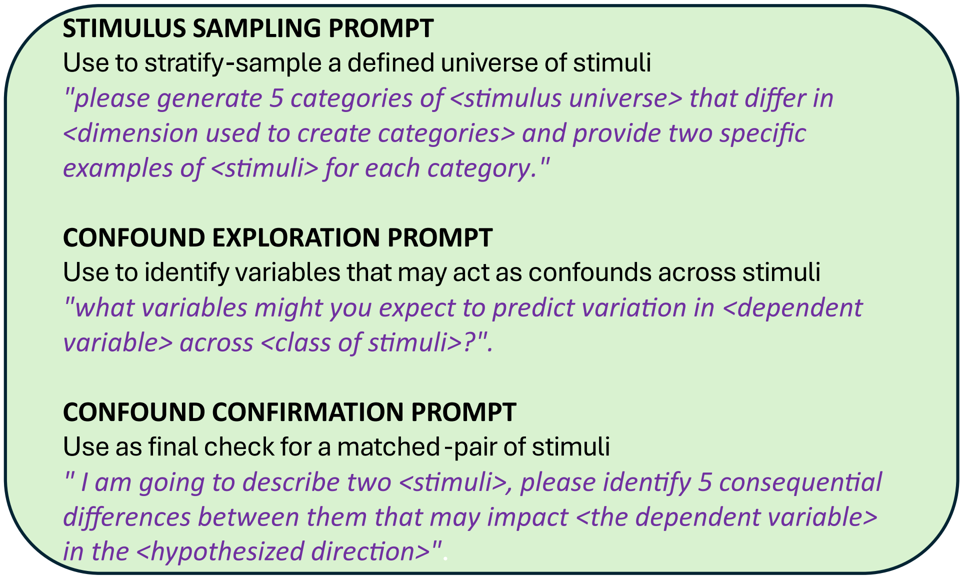
Scenario 1 <copy pasted full scenario with chopping vegetables>

Scenario 2 <copy pasted full scenario with self-harm>".

The five variables that were identified by ChatGPT were (i) emotional benefit to John, (ii) urgency of the need, (iii) concern for John's mental health, (iv) sense of social responsibility (for John), and (v) increase awareness of mental health issues more generally. With this feedback it seems straightforward to iterate and modify the scenario to reduce potential confounds (in Supplement 10 we report results of this confound confirmation prompt for all 20 stimuli in the study).

It's worth noting that some confounds are subtle and hard to detect, and are likely to be missed by GenAI tools or participants. Therefore, it's advisable to use this prompt as a complement to, rather than a substitute for, careful expert judgment. We are not delegating to GenAI this task, we are using GenAI as an assistant. Figure 7 summarizes the different prompts that can be posed to hypothesis-blind agents. This footnote explains why we think the first two prompts are more reliable.[16] Figure 7 contains a flowchart summarizing Mix-and-Match.

[16] At the time of writing, we suspect that GenAI tools are better suited for stratified-sampling of stimuli and identifying potential confounds than for ruling out confounds in specific stimulus-pairs. This is in part because GenAI tools are particularly good at organizing existing information (e.g., all that is known about what predicts word recognition, or

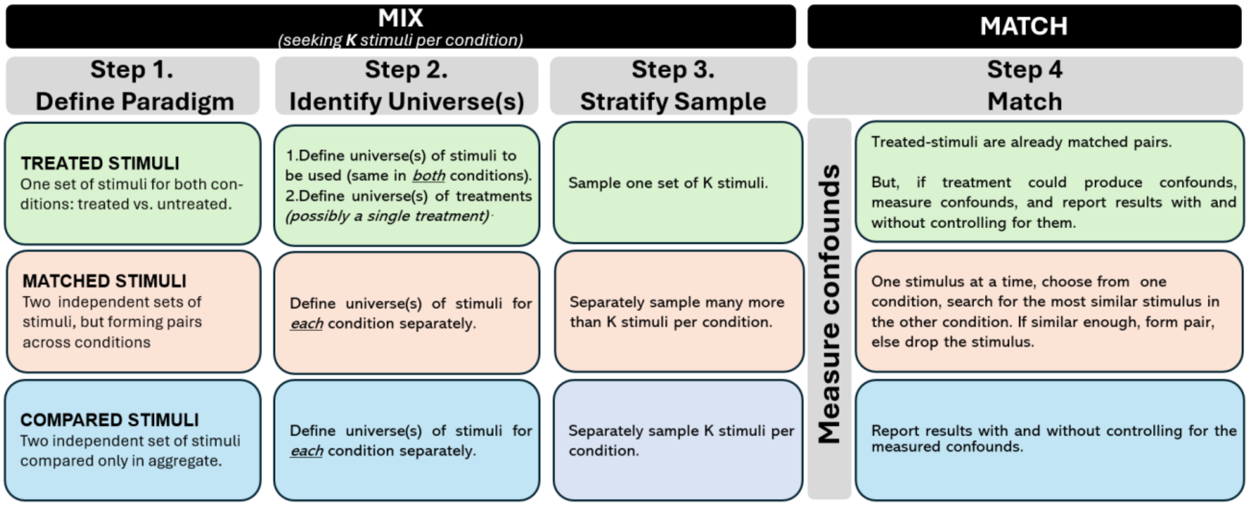


STIMULUS SAMPLING PROMPT
Use to stratify-sample a defined universe of stimuli
"please generate 5 categories of <stimulus universe> that differ in <dimension used to create categories> and provide two specific examples of <stimuli> for each category."

CONFOUND EXPLORATION PROMPT
Use to identify variables that may act as confounds across stimuli
"what variables might you expect to predict variation in <dependent variable> across <class of stimuli>?".

CONFOUND CONFIRMATION PROMPT
Use as final check for a matched -pair of stimuli
" I am going to describe two <stimuli>, please identify 5 consequential differences between them that may impact <the dependent variable> in the <hypothesized direction>".

Figure 7. *Three Key Questions to Pose to Hypothesis-Blind Agents for Mix-and-Match*



MIX *(seeking **K** stimuli per condition)*				MATCH
Step 1. Define Paradigm	**Step 2. Identify Universe(s)**	**Step 3. Stratify Sample**		**Step 4 Match**
TREATED STIMULI One set of stimuli for both conditions: treated vs. untreated.	1.Define universe(s) of stimuli to be used (same in ***both*** conditions). 2.Define universe(s) of treatments *(possibly a single treatment)*	Sample one set of K stimuli.	**Measure confounds**	Treated-stimuli are already matched pairs. But, if treatment could produce confounds, measure confounds, and report results with and without controlling for them.
MATCHED STIMULI Two independent sets of stimuli, but forming pairs across conditions	Define universe(s) of stimuli for ***each*** condition separately.	Separately sample many more than K stimuli per condition.		One stimulus at a time, choose from one condition, search for the most similar stimulus in the other condition. If similar enough, form pair, else drop the stimulus.
COMPARED STIMULI Two independent set of stimuli compared only in aggregate.	Define universe(s) of stimuli for ***each*** condition separately.	Separately sample K stimuli per condition.		Report results with and without controlling for the measured confounds.

Figure 8. Overview of Mix-and-Match

In Supplements 1-4 we apply Mix-and-Match to four different study designs from published papers. The classic Asian Disease problem by Tversky and Kahneman (1981), and the three studies we re-analyzed using Stimulus Plots.

sources of embarrassment) but less good at bringing up absent contextual & background information on specific instances. In our own experience, GenAI often surprised us with excellent suggestions for strata, performed well identifying candidate confounds and was only OK identifying confounds in specific stimuli pairs, often raising secondary irrelevant aspects and often missing confounds that would be more obviously evidence to a human evaluator. For examples see Supplement 10.

The future of stimulus sampling

We envision a future where running multi-stimuli experiments becomes the norm in psychological research. In this future, researchers prioritize treated-stimulus designs whenever possible and carefully select stimuli from stratified samples using 'stimulus sampling prompts'. They also recognize that matched-stimulus designs require particular attention to confounds, taking proactive steps to rule them out. A future where researchers follow and document the steps in Mix-and-Match (see our Appendix A for a proposed disclosure form), making it straightforward for peer-reviewers and readers to evaluate design choices and consider principled variations of those choices.

In this ideal future, experimental results are always presented at the individual stimulus level using Stimulus Plots. When these reveal substantive heterogeneity, authors attempt to disentangle confounds from moderators as likely explanations, reporting results from "confound confirmation prompts". When plausible confounders are identified, aggregate results without the suspected stimuli are reported for robustness, and future studies in the same project address or remove the suspicious stimuli. When it comes to evaluating existing work which did not follow the stimulus sampling ideal we imagine here, we believe that retroactively creating Stimulus Plots and running confound confirmation prompts can help re-interpret past findings and improve the design of future studies.

General Discussion

We close by touching on a series of issues we expect readers may be thinking about as they reach this last section of the paper.

Isn't e*xternal validity also important?* Prior papers on the selection and analysis of experiments with multiple stimuli have focused on external validity. We have already argued in detail why the emphasis should instead be on internal validity. But to be clear, we do believe external validity is valuable. If something only happens in contrived lab environments, it is not clear psychologists should care about it, and in any case they should be aware that it does only happen in contrived lab environments. However, we don't think that external validity involves testing different stimuli (which may or may not be internally valid) within the same paradigm. Rather, external validity for an experimental paradigm can only be assessed by collecting data *outside* that paradigm; and to know that a finding is consequential in the real world, a perhaps more common understanding of *external* validity, the findings need to be documented… …in the real world.

Does using multiple stimuli reduce statistical power? One concern we believe people may have with our call for routinely using multiple stimuli in experiments, is that doing so may lower power to detect an overall effect. This concern is largely unfounded. Adding stimuli could indeed decrease power if authors knew which stimulus shows the largest effect and were to choose, in the absence of stimulus sampling, that single stimulus for their experiment. But, a more likely scenario is that experimenters don't know for sure which stimuli will show larger effects, and in that case adding stimuli will tend to increase power.[17] Additionally, if one is able to present more than one stimulus per participant, adding stimuli also increases power. Thus, power concerns, if anything, generally provide additional justification for multiple stimuli.

[17] To get an intuition for this: imagine two stimuli, one has a very big effect detectable with any sensible sample size, the other no effect. Using only one of them, blindly, expected power is 52.5% (since 105%/2 = 52.5%). If the study uses both, instead, the one that is very big will make the entire study 'work', power of 100%.

Why within a study? An interesting question we have received is 'what is the benefit of running one study with many stimuli instead of many studies with one stimulus each?' First, running multiple stimuli with a given paradigm in one study, allows changing the paradigm across studies, which is valuable for internal and perhaps external validity. Second, running multiple stimuli in the same study allows differences in results across stimuli to be causally interpretable (as they arise under random assignment and/or from the same participants). Third, transparent reporting of all stimuli attempted is verifiable if done in one study (that's pre-registered), but not across studies (which may be file-drawered). Fourth, researchers often rely on the fallacious argument that if each study in a paper suffers from a different confound then the 'parsimonious' explanation is the one of interest to the authors as if it is the only one that account for all the data (Simonsohn, 2014). Having all stimuli in one study precludes this fallacious way of thinking about confounds and parsimony during the design and analysis of studies. Fifth, as mentioned above, multiple stimuli in a study can increase power.

Isn't the implementation of Mix-and-Match subjective and arbitrary? In short. Yes. But… It is *less* subjective and *less* arbitrary than the status quo where researchers follow undisclosed and presumably unsystematic procedures of stimulus selection. Mix-and-Match does not eliminate idiosyncrasies in how psychologists operationalize hypotheses, but it reduces those idiosyncrasies, it highlights them, and it provides a framework for discussing them.

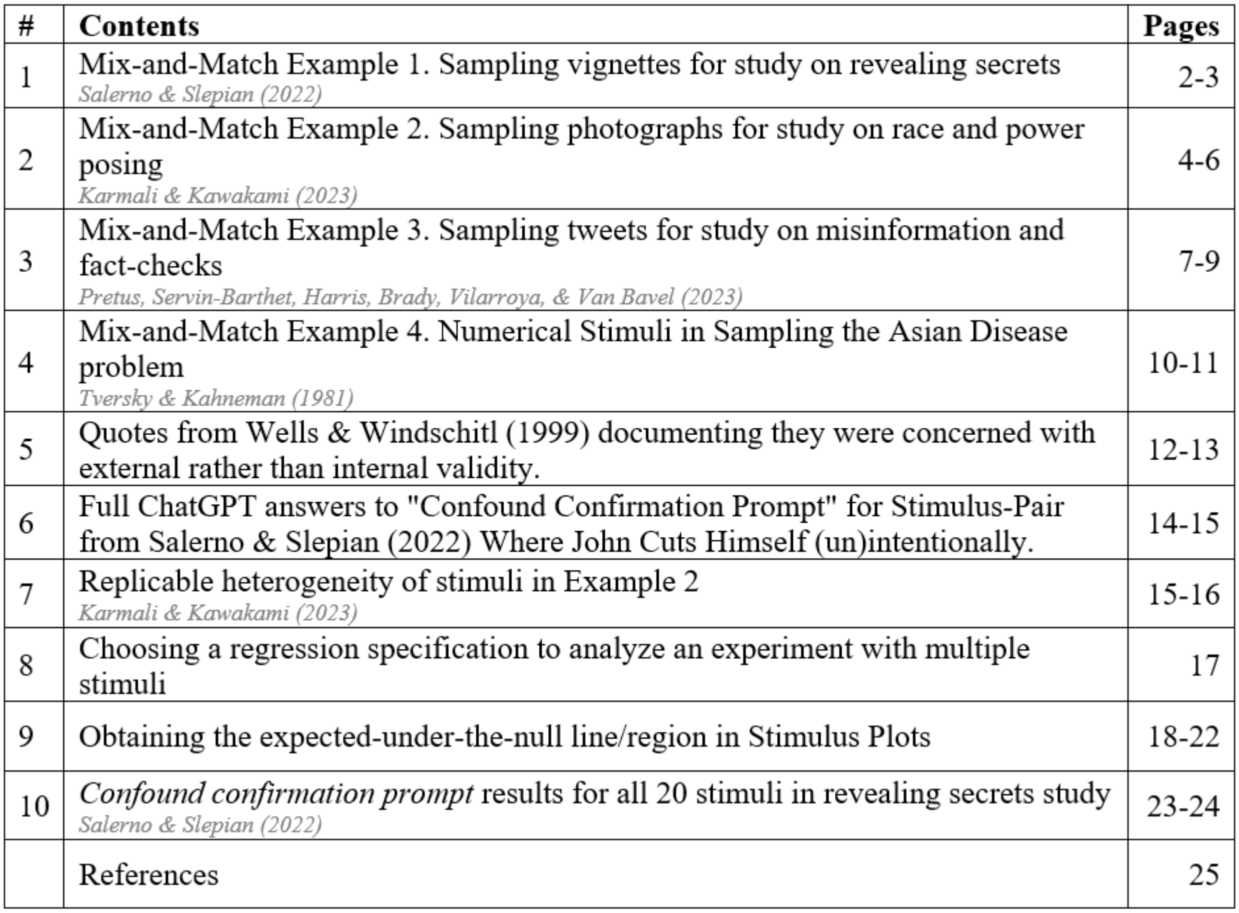
Doesn't mediation take care of internal validity? The goal of mediation is indeed to ascertain whether a randomly assigned manipulation produces an observed effect through a hypothesized channel. But, it has long been recognized that mediation analysis does not deliver on its stated goal (Bullock & Green, 2021; Bullock, Green, & Ha, 2010; Judd & Kenny, 1981, pp. 607, last paragraph; Rohrer, Hünermund, Arslan, & Elson, 2022). Most notably, mediation analysis is biased towards finding mediation which does not exist under two likely scenarios. First, if the mediator is correlated with the dependent variable outside of the experiment (for the intuition, see Simonsohn, 2022), and second, if the stimuli across conditions differ in more than in the attribute of interest and those alternative mediators are not included in the analysis.

Limitations. In this paper we have proposed new tools, and all tools from pencils to rearview mirrors, can be misused. We discuss some potential misuses, hoping readers will avoid them. For Mix-and-Match, a possible misuse involves mixing and/or matching over superficial dimensions, leading to studies that do not include truly diverse stimuli or fail to match stimuli on the relevant confounds. We hope that our proposed Mix-and-Match disclosure form, see Appendix A, which describes the step-by-step procedure, will help authors avoid these issues and assist readers in evaluating implementations of Mix-and-Match.

For Stimulus Plots, a possible misuse involves unreasonably expecting all stimuli to conform to predictions, be it with authors file-drawering results because some stimuli do not behave as expected, or reviewers encouraging authors to "explain" something they cannot really explain. We hope the confidence band we include in Stimulus Plot, and the disclaimers we have offered throughout the article will be effective protection against such misuse.

In addition to potential misuse, a limitation of our proposals is that, while we strived to make them broadly applicable across psychology, our personal expertise and experience is likely to pose blind spots to challenges in applying our recommendations to fields that are further from our own (judgment and decision-making). For example, how to apply the stimulus sampling prompt to perception research, or to in person studies where participants interact with a confederate, needs to be worked out by colleagues with relevant expertise. Implementation aside, we believe the recommendations in this article apply to any behavioral experiment where it is relevant to understand why the chosen stimuli show the effect that they do.

We close with the table of contents to the supplementary materials, and drawing attention to all readers to Appendix A which includes a Mix-and-Match Disclosure Form to be included in papers that reports experiments designed relying on Mix-and-Match.



#	Contents	Pages
1	Mix-and-Match Example 1. Sampling vignettes for study on revealing secrets *Salerno & Slepian (2022)*	2-3
2	Mix-and-Match Example 2. Sampling photographs for study on race and power posing *Karmali & Kawakami (2023)*	4-6
3	Mix-and-Match Example 3. Sampling tweets for study on misinformation and fact-checks *Pretus, Servin-Barthet, Harris, Brady, Vilarroya, & Van Bavel (2023)*	7-9
4	Mix-and-Match Example 4. Numerical Stimuli in Sampling the Asian Disease problem *Tversky & Kahneman (1981)*	10-11
5	Quotes from Wells & Windschitl (1999) documenting they were concerned with external rather than internal validity.	12-13
6	Full ChatGPT answers to "Confound Confirmation Prompt" for Stimulus-Pair from Salerno & Slepian (2022) Where John Cuts Himself (un)intentionally.	14-15
7	Replicable heterogeneity of stimuli in Example 2 *Karmali & Kawakami (2023)*	15-16
8	Choosing a regression specification to analyze an experiment with multiple stimuli	17
9	Obtaining the expected-under-the-null line/region in Stimulus Plots	18-22
10	*Confound confirmation prompt* results for all 20 stimuli in revealing secrets study *Salerno & Slepian (2022)*	23-24
	References	25

Table 1. Contents of supplementary materials.
Available from https://researchbox.org/2257/47 (use code: CXUWHS)

References

Bar-Hillel, M., Maharshak, A., Moshinsky, A., & Nofech, R. (2012). A rose by any other name: A social-cognitive perspective on poets and poetry. *Judgment and Decision making, 7*(2), 149-164.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., . . . Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences, 115*(11), 2607-2612.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological review, 62*(3), 193.

Bullock, J. G., & Green, D. P. (2021). The failings of conventional mediation analysis and a design-based alternative. *Advances in Methods and Practices in Psychological Science*, *4*(4), 25152459211047227.

Bullock, J. G., Green, D. P., & Ha, S. (2010). Yes, But What's the Mechanism?(Don't Expect an Easy Answer). *Journal of personality and social psychology, 98*(4), 550-558.

Campbell, D. T., & Cook, T. D. (1979). Quasi-experimentation. *Chicago, IL: Rand Mc-Nally, 1*(1), 1-384.

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*(4), 335-359.

Dias, N., & Lelkes, Y. (2022). The nature of affective polarization: Disentangling policy disagreement from partisan identity. *American Journal of Political Science, 66*(3), 775-790.

Evangelidis, I., Levav, J., & Simonson, I. (2023). The upscaling effect: how the decision context influences tradeoffs between desirability and feasibility. *Journal of Consumer Research, 50*(3), 492-509.

Haidt, J., McCauley, C., & Rozin, P. (1994). Individual differences in sensitivity to disgust: A scale sampling seven domains of disgust elicitors. *Personality and Individual differences*, *16*(5), 701-713.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation review, 5*(5), 602-619.

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*(1), 54.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual review of psychology*, *68*(1), 601-625.

Karmali, F., & Kawakami, K. (2023). Posing while black: The impact of race and expansive poses on trait attributions, professional evaluations, and interpersonal relations. *Journal of personality and social psychology*, *124*(1), 49-68.

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science, 10*(4), 518-536.

Lerner, J., Small, D. A., & Loewenstein, G. F. (2004). Heart strings and purse strings - Carryover effects of emotions on economic decisions. *Psychological science, 15*(5), 337-341.

McNeish, D. (2023). A practical guide to selecting and blending approaches for clustered data: Clustered errors, multilevel models, and fixed-effect models. *Psychological methods*.

Novoa, G., Echelbarger, M., Gelman, A., & Gelman, S. A. (2023). Generically partisan: Polarization in political communication. *Proceedings of the National Academy of Sciences, 120*(47), e2309361120.

Pretus, C., Servin-Barthet, C., Harris, E. A., Brady, W. J., Vilarroya, O., & Van Bavel, J. J. (2023). The role of political devotion in sharing partisan misinformation and resistance to fact-checking. *Journal of Experimental Psychology: General*.

Rohrer, J. M., Hünermund, P., Arslan, R. C., & Elson, M. (2022). That's a lot to PROCESS! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221095827.

Rosenthal, R. (2009). Blind and Minimized Contact. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifacts in Behavioral Research* (pp. 592-602): Oxford University Press.

Rubenstein, H., Lewis, S. S., & Rubenstein, M. A. (1971). Evidence for phonemic recoding in visual word recognition. *Journal of verbal learning and verbal behavior, 10*(6), 645-657.

Salerno, J. M., & Slepian, M. L. (2022). Morality, punishment, and revealing other people's secrets. *Journal of personality and social psychology, 122*(4), 606.

Simonsoh, U. M., Andres; Evangelidis, Ioannis. (2024). Stimulus Sampling Reimagined Retrieved from https://researchbox.org/2257

Simonsohn, U. (2014). [31] Women are taller than men: Misuing Occam's Razor to lobotomize discussions of alternative explanations. Retrieved from https://datacolada.org/31

Simonsohn, U. (2022). [103] Mediation Analysis is Counterintuitively Invalid. Retrieved from https://datacolada.org/103

Spiller, S. A. (in press). Commentary on Eskreis-Winkler and Fishbach (2019): A Tendency to Answer Consistently Can Generate Apparent Failures to Learn From Failure. *Psychological science*.

Strickland, B., & Suben, A. (2012). Experimenter philosophy: The problem of experimenter bias in experimental philosophy. *Review of Philosophy and Psychology, 3*, 457-467.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

Wells, G., & Windschitl, P. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25*(9), 1115.

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. *PloS one, 11*(3), e0152719. doi:10.1371/journal.pone.0152719

Appendix A: Mix-and-Match disclosure form

We encourage authors who rely on Mix-and-Match to include this form in a supplement to their paper to report in a standardized fashion how they designed their study.

Mix-and-Match Disclosure Form

Step 1: Define Paradigm.
Instructions: Provide a clear definition of the experimental paradigm, specifying whether a treated, matched, or compared-stimulus design is used, and describe the dependent variable.

Example:
A 2-cell matched-stimulus design, where participants are presented with words that either have or do not have a homophone. The dependent variable is whether participants recognize the word as valid (yes/no).

Step 2: Identify Universe(s).
Instructions: Describe the universe(s) of stimuli for the chosen paradigm, outlining the relevant categories from which stimuli will be sampled.

Example:
Universe 1 (categorical): All words with a homophone in the Oxford English Dictionary.
Universe 2 (categorical): All words without a homophone in the Oxford English Dictionary.

Step 3. Stratify Sample.
Instructions: Enter the Stimulus Sampling Prompt and the resulting strata and stimuli.

Example:
We submitted this Stimulus Sampling Prompt to ChatGPT: "Please generate 5 categories of homophones that differ in their etymological origin,and provide two examples of specific homophones for each."

Stratum 1: Latin vs. Germanic Origins (Alter/Altar, Peace/Piece)
Stratum 2: Old French vs. Old English Origins (Pair/Pear, Scent/Sent)
Stratum 3: Greek vs. Old English Origins (See/Sea, Cell/Sell)
Stratum 4: Old Norse vs. Old English Origins (Flower/Flour, Fowl/Foul)
Stratum 5: Dutch vs. Latin Origins (Right/Rite, Vein/Vain)

Step 4. Match.
Instructions: Explain how you ensured that the stimuli across conditions differ only on the focal attribute of interest. For matched- and compared-stimulus design include a confound exploration prompt and the results.

Example.
We submitted this Confound Exploration Prompt to ChatGPT: "What 5 variables might be expected to predict variation in reaction time to recognize a word as valid, across different words?", identifying Word Frequency, Word Length, Word Familiarity, Phonological Complexity, and Semantic Concreteness.

We used a linguistic database containing ratings of words across these identified confounds to find the non-homophone word that is most similar to each homophone word identified in Step 3.