This version: 2024 06 27
Newest version: 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 latent variables that cannot be directly manipulated (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 results from psychology experiments. This, however, is generally not the case, at least not for the causal claims of interest to psychologists. The reason is that, in contrast to the hard and some applied sciences, where experimenters can directly manipulate the independent variable of interest (e.g., physicists can directly manipulate an object's mass, economists can directly manipulate what the default option is on a tax form), many psychology experiments examine hypotheses about latent variables that are neither observable nor directly manipulable (e.g., trust, self-worth, impatience, risk-tolerance, arousal). Psychology researchers, therefore, typically indirectly manipulate variables of interest by randomly assigning participants to conditions with different stimuli. Given that stimuli are multidimensional, any two stimuli that participants are randomly assigned to will typically differ not only on the focal dimension the experimenter wishes to manipulate (e.g., the emotional reaction they induce), but also on other dimensions the experimenter does not wish to manipulate. In psychology we randomly assign stimuli to participants, but we seldom randomly assign attributes to stimuli.

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.

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

[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 Brunswik (1955) and the rest of the special issue published in *Psychological Review* V62(3).

[2] Wells and Windschitl (1999) write that "failure to sample stimuli also can threaten ***construct validity***." (emphasis added; their abstract). But as we document in Supplement 5, all arguments in their article (except for their footnote 9), involve external rather than construct validity.

[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.

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.

Let's start by addressing 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 is the intended one, then one stimulus per condition would be enough for internal validity purposes. In practice, however, we never know with high enough confidence that the only difference between any given pair of stimuli is the intended one. 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 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 stimulus samples can be too small for internal validity purposes, they can also be too large. The reason is that as the sample of stimuli grows, so does the amount of random variation among them in any given sample, obfuscating 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]

Thus, once a study has 5 or 10 stimuli per condition there is 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 psychologists do not run studies to learn about the properties *of the stimuli* they use, they run studies to learn about *people*. Stimuli are 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.

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,

[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 explored 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.

[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.

which they obviously do, the disgust manipulation would be confounded. Again, we randomly assign participants to watch a clip, we don't randomly assign disgust to a given movie scene. 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.

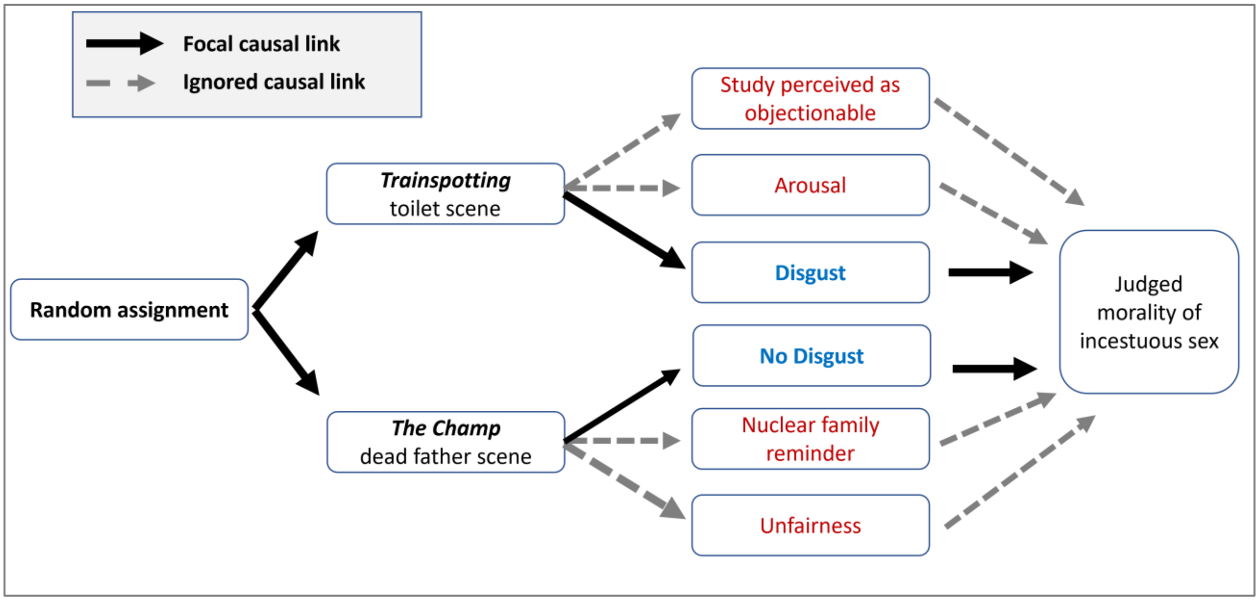

Focal causal link
Ignored causal link
Random assignment
Trainspotting toilet scene
The Champ dead father scene
Study perceived as objectionable
Arousal
Disgust
No Disgust
Nuclear family reminder
Unfairness
Judged morality of incestuous sex

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.

In this paper, therefore, 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 generating stimuli: 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 and documentable process of stimuli generation which helps researchers be transparent about how and why they operationalize their latent constructs 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 analysis: 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). Our proposal of constructing "Stimulus Plots" changes all of this.[6]

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.

[6] The output of a mixed model *can* be used to explore variation, but it is not standard or easy to do so: In R the estimates for each stimulus can be revealed with the command lme4::ranef().

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. In what follows we begin with our proposed analysis of multi-stimuli studies, and then proceed to their design.

Stimulus Plots

Only by analyzing data at the individual stimulus level can the main goal of stimulus sampling be achieved: assessing internal validity.[7] 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

[7] 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.

in the opposite direction? Are there outlier stimuli with surprisingly big or small effects that may shed light on confounds or moderators?

We propose analyzing individual stimuli relying on what we refer to as "Stimulus Plots", plotting stimuli-level results side-by-side, sorted by effect size. Stimulus Plots have two panels: one plots the means by condition, the other differences of means, the effects, across conditions (note: proportions are also means). While these plots are exploratory, we propose also 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').[8] 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 will provide an R package, 'stimulus', with a function that makes Stimulus Plots in one line of code.

We next illustrate 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, and although he had decided beforehand, he would not take any illegal party drugs,*

[8] The approach to obtaining the expected level of variation across stimuli if the true effects were identical involves resampling by shuffling the column with the stimulus ID, see Simonsohn, Simmons, and Nelson (2020).

a friend offered him some, and in the heat of the moment, he said yes." (see their Appendix C; p.24).[9]

The authors report 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. We obtained their posted data and reproduced this result. Then we tested for heterogeneity in effect size across stimuli through an ANOVA model comparison test (see e.g., McNeish, 2023), obtaining statistically significant results, $\chi^2(2)=74.24$, $p < .0001$.[10] The Stimulus Plots in Figure 2 allows us to understand the nature and implications of this heterogeneity.

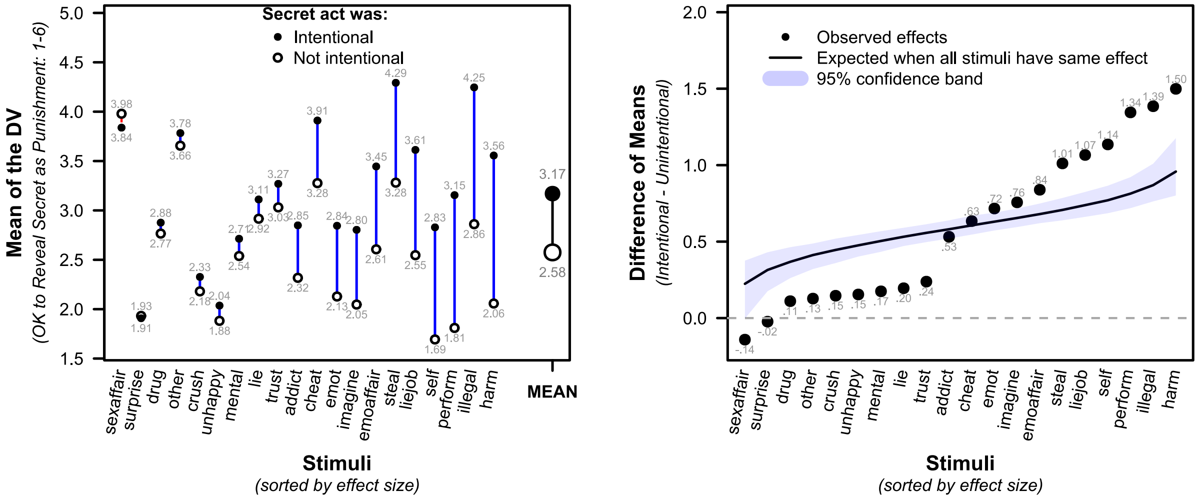


Figure 2. Stimulus Plots for Study 4 in Salerno & Slepian (2022)

The study involves a 2-cell treated-stimuli 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 stimuli after shuffling the stimulus label across rows repeatedly.

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

The left panel shows that about 7 stimuli exhibit no difference across conditions, while several stimuli show very large differences. The right panel contrasts this variation with what

[9] The word "and" appears twice in a row in the original text we quote from.

[10] To test for heterogeneity we estimated a mixed model which included only random intercepts for stimuli, and one with also random slopes for stimuli, and did an ANOVA model comparison of the two. In R syntax:

```
homo  <- lme4::lmer(rev~intent+(1|participant_id)+(1|stimulus))
hetero <- lme4::lmer(rev~intent+(1|participant_id)+(intent|stimulus))
anova(homo, hetero)
```

would be expected if all stimuli were equally effective, and all variations were due to sampling error. The data exhibit much more heterogeneity than expected. This expected line is constructed by bootstrapping (under the null), specifically, by repeatedly shuffling the column in the data with the stimulus identifier (this is analogous to how Specification Curves are computed under the null, see Simonsohn et al. (2020)).

Contrasting expected with observed variation, we see the data have seven stimuli with effect sizes smaller than what we would expect the smallest effect to be, and six that are larger than what the largest effect would be expected to be.

That a substantial share of stimuli "do not work" in this study does *not* necessarily invalidate its main conclusion, but does warrant a deeper exploration of the design and results than is provided in the article. 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"). See Appendix C in Salerno & Slepian 2022, p.24. 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. It is speculative of course, whether that's why that stimulus shows such a large effect. 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.

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.[11] We begin with Study 3, as its Stimulus Plot revealed important information that was left undetected and unexplored in the original paper.

In the study, n=105 undergraduates were asked to choose potential partners for an upcoming task. They saw 20 sets of 4 photographs of different people, and they chose one out of the four in each set as a potential partner. 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 and reproduced this result. Then we tested for heterogeneity in effect size across stimuli through an ANOVA model comparison test (see e.g., McNeish, 2023), obtaining statistically significant results both among the 20 Black potential partners, $\chi^2(2)=94.24$, $p < .0001$, and 20 White potential partners, $\chi^2(2) = 19.74$, p<.0001. The Stimulus Plots in Figures 3 and 4 allow us to understand the nature and implications of this heterogeneity.

We observe that while *on average* Black potential partners are not more or less likely to be chosen in expansive rather than constrictive poses, posing has a highly heterogeneous effect. There are eight Black potential partners who exhibit a *negative* effect of expansive posing, seven of which show an effect *bigger* in magnitude than the average (positive) effect for White potential partners. There are, however, also several Black potential partners showing strong effects in the oppositive direction, cancelling out on average. Note that the biggest effect is a whopping 38 percentage point increase in the probability of being chosen.

[11] The design involves 5 expansive and 5 contractive poses. Any given potential target was shown in 2 out of 5 poses of each kind.

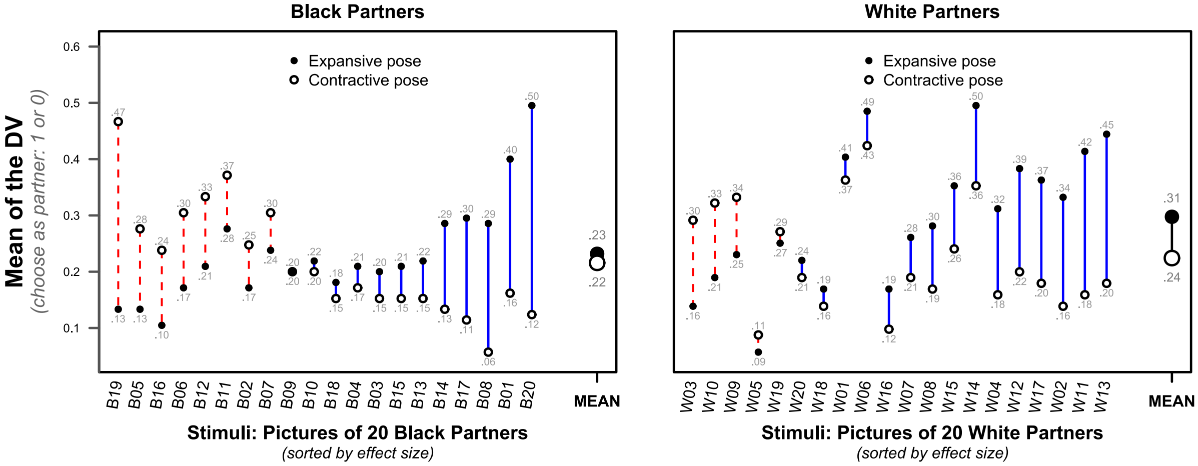


Figure 3. Stimulus Plot for Means in Karmali & Kawakami – Study 3
The study involves a 2 (race [compared]) x 2 (power posing [treated]) stimuli design. Participants chose 1 of 4 potential partners based on photographs where they were either in an expansive or a contractive pose. The figure depicts the percentage of times each stimulus (potential partner) was chosen.
R Code to reproduce the figure: https://researchbox.org/2257/49 (use code CXUWHS)

Figure 4 reports differences of means for each stimulus, contrasting observed differences with what be expected under homogeneity. We believe the Stimulus Plots from Figures 3 and 4 make clear that there is important heterogeneity to explore before interpreting the results from this study in the way they have been interpreted.

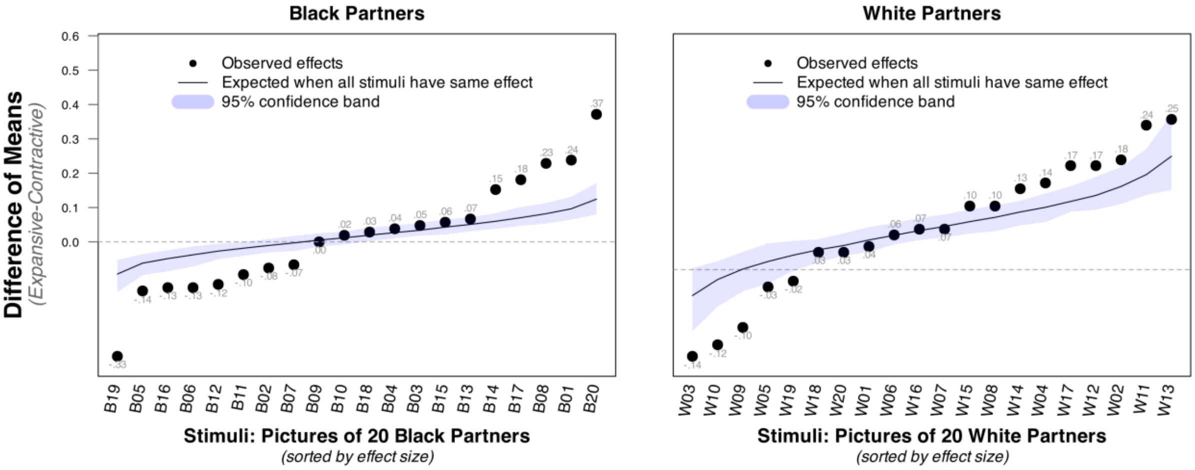


Figure 4. Stimulus Plot for Effects in Karmali & Kawakami – Study 3
Differences computed off means from Figure 5. The expected line, and its 95% confidence interval, are obtained via resampling.
R Code to reproduce the figure: https://researchbox.org/2257/49 (use code CXUWHS)

Indeed, if we focus on average *magnitude* of the effect, that is, if we compute the mean absolute effect, Black targets were directionally *more* influenced on average by the pose, with an average effect of 12.7 percentage points, compared to 11.0 percentage points for White targets. Thus, the summary conclusion by the original authors that "expansive versus constrictive poses did not influence participants' willingness to interact with Black targets" (p.59) seems contradicted by their data once individual, rather than 'average' stimuli, are analyzed. Perhaps most importantly, the variation we see among Black targets is so substantial that it suggests the presence of confounds (e.g., the same person in an expansive vs contractive pose looked less inviting as a potential partner for reason other than the pose itself), or perhaps moderators (e.g., expansive poses make targets appealing, but only if they are smiling or have short hair). But, keep in mind that the Stimulus Plots would imply a moderator that is not evenly distributed across White and Black targets, and so, even a moderator explanation is ultimately a confound explanation for the key contrast of interest: posing for Blacks vs Whites.

Understanding exactly what's behind the heterogeneity we uncovered seems worthwhile, but it requires access to the original stimuli (the photos). The authors did not post them and did not provide them to us despite a few requests.

While we have attempted to guard against chasing noise in exploring heterogeneity, through formal heterogeneity tests, and by including a confidence band around the expected variation across stimuli under homogeneity, there is a final check we can do for this particular example because the original paper conducted a replication of the study we re-analyzed. Specifically, Study 4 by Karmali & Kawakami (2023) presented the same stimuli and collected the same dependent variable from a new set of participants. Following a suggestion by the second author of that paper, we contrasted the stimulus level estimates across studies to assess how stable the differences we documented were. The results are shown on Figure 5. The heterogeneity is extremely stable; the correlation of effects by stimulus across the two studies is r = .85, p < .0001. Future research that builds on this work should strive to identify the confounds or moderators behind this replicable heterogeneity.

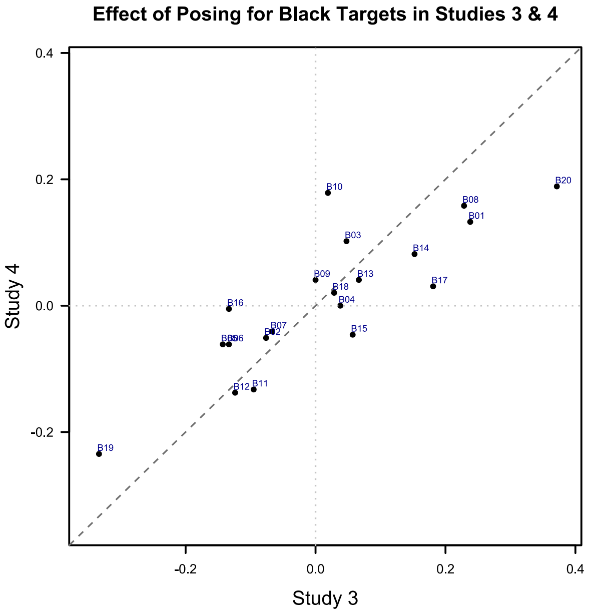


Figure 5. The Same Stimuli Lead to Heterogeneous but Replicable Results in Studies 3 & 4
Original data collected by Karmali & Kawakami (2023). Each dot represents the effect of being in an expansive vs contractive pose for Black potential targets. For example, potential Black partner "B19" was 33% less likely to be chosen in the expansive vs contractive pose in Study 3, and 23% less likely in Study 4. The overall correlation in effect within stimulus across experiments is r = .85, p<.0001.

Example 3. All stimuli are 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). The paper reports an overall average effect of the fact-check of M=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 tested for heterogeneity in effect size across stimuli through an ANOVA model comparison test (see e.g., McNeish, 2023), obtaining statistically non-significant results, $\chi^2(2)=0.933$, $p = .627$. We report Stimulus Plots for this study in Figure 6. The left panel shows some variation in effect size across stimuli, but the right panel shows that the observed level of variation is consistent with sampling error; this is consistent with the heterogeneity test just reported.

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 the there is no heterogeneity, but rather, that to study heterogeneity for these stimuli one needs a larger sample of participants.

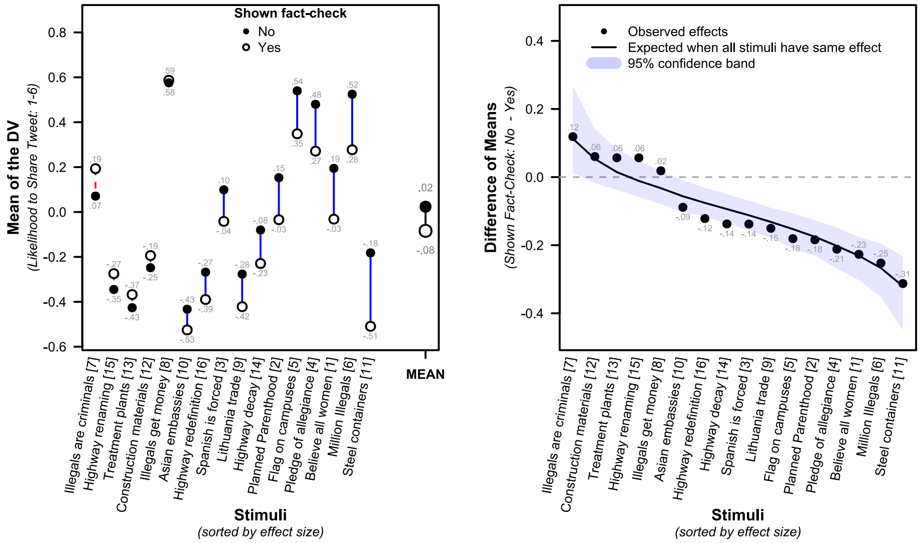


Figure 6. Stimulus Plots for Pretus et al. (2023) – Study 2
The study involves a two-cell treated-stimuli design, comparing participants' reported willingness to share a tweet containing information having been presented, or not, with a Twitter fact-check. The expected line, and its 95% confidence interval, 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).

Stimulus Plots Contribution to the Statistical Analyses of Clustered Data

To appreciate how Stimulus Plots upend the conclusions we may draw from traditional statistical analyses of experimental results, in this section we carry out such analyses for our three examples and contrast the conclusions with vs without considering Stimulus Plots.

Data from experiments with multiple stimuli are almost always 'clustered': a given stimulus is shown to multiple participants, and multiple observations are often collected from individual participants. Clustered data are commonly analyzed in one of two ways: with (1) regressions that include fixed-effects, clustered standard errors, or both, or with (2) mixed-models with random intercepts and possibly random slopes. McNeish (2023) provides a detailed overview of both approaches and how they relate to each other and to the research question motivating data collection. We conducted our analyses with both approaches.

Estimated models

We estimated three models for each dataset. First, we estimated a regression model with stimuli and participant (ID) fixed effects, i.e., 1/0 dummies for each ID, to control for variability across them (this increases power), and clustering errors by participant, to account for any lack of independence that may remain across observations by the same ID beyond their fixed effect.

In R Code: `miceadds::lm.cluster(dependent variable~condition+factor(stimulus)+factor(ID), cluster=ID)`

Second, we estimated a mixed-effects model that included random intercepts for participants and for stimuli. In a mixed-model these intercept increase power by controlling for stimuli heterogeneity and (partially) address dependence across multiple observations by the same participant.[12]

In R Code: `lme4::lmer(dependent variable ~ condition + (1|stimulus) + (1|ID))`

Third, we estimated a mixed-effects model that includes also 'random slopes' for stimuli. These 'random slopes' correspond to the effect of the manipulation for each stimulus (e.g., what the effect of intentionality was on the tendency to divulge a secret for each of the vignettes).

In R Code: `lme4::lmer(dependent variable ~ condition + (1+condition|stimulus) + (1|ID))`

A mixed-model that includes random slopes typically enlarges the confidence interval for the main effect of the manipulation based on the variability of the estimated effect across stimuli, the standard deviation across slopes (e.g., the confidence interval around the overall estimate on divulging secrets is increased by the variability in the observed effects across the 20 vignettes). Including random slopes, then, usually lowers statistical power.

It has been proposed that (these power lowering) random slopes for stimuli should be used whenever the data allow, in order to generalize findings obtained with some stimuli to results that could be obtained with other stimuli, and that without random slopes the overall test of the manipulation has an elevated false-positive rate (see e.g., Barr, Levy, Scheepers, & Tily, 2013;

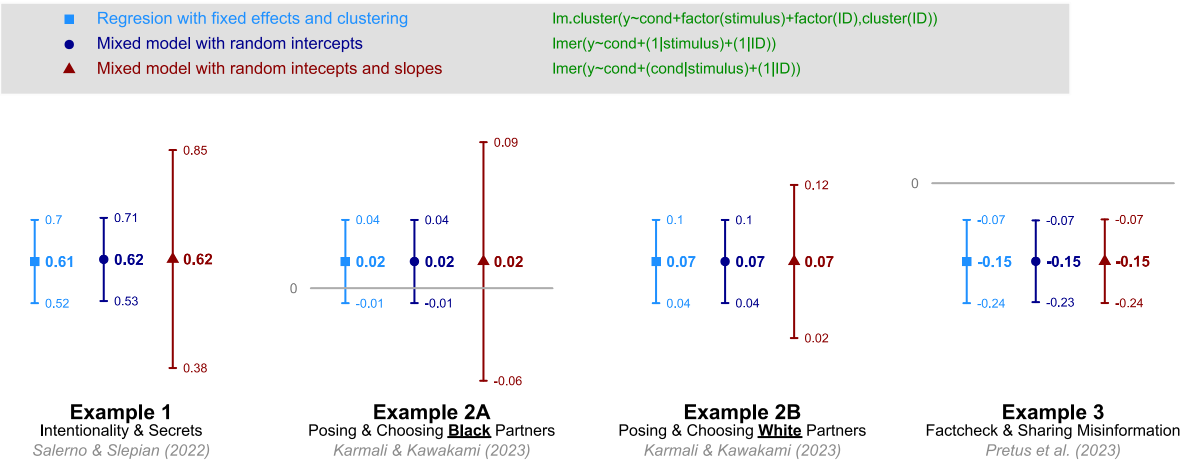
[12] We say "partially" accounting for dependence because a random intercept only accounts for dependence that arises from different participants having different mean evaluations. It's equivalent to the fixed effect but the mixed model lacks the catch-all ability to account for *other* dependency that the regression has with clustered errors (Abadie, Athey, Imbens, & Wooldridge, 2023; McNeish, 2023).

Brauer & Curtin, 2018; Judd et al., 2012; Oberauer, 2022; Wickens & Keppel, 1983). We believe these claims rest on rather unrealistic assumptions and we are not persuaded random slopes are necessary for valid inferences. But we report results with and without random slopes, readers can individually decide which to focus on.

Results

Figure 7 shows the results of these three alternative models estimated on each of the datasets discussed above. First, note that for any given dataset, the point estimates of the average effect of the manipulation are essentially identical across the three procedures. Second, the confidence intervals are essentially identical for the regression and the mixed model with random intercepts.[13] Third, in Examples 1, 2a and 2b, the confidence interval gets meaningfully wider upon including random slopes, this occurs because there is meaningful heterogeneity in the effect size across stimuli for them. As mentioned above, random-slopes models add such heterogeneity to the confidence interval of the overall effect, hence lowering power. Because in Example 3 there is no more heterogeneity than expected by chance, the random slopes leave the confidence interval unchanged.

[13] This implies there is no consequential dependence within participants after accounting for their mean response (their fixed/random effect).



```
lm.cluster(y~cond+factor(stimulus)+factor(ID),cluster(ID))
lmer(y~cond+(1|stimulus)+(1|ID))
lmer(y~cond+(cond|stimulus)+(1|ID))
```

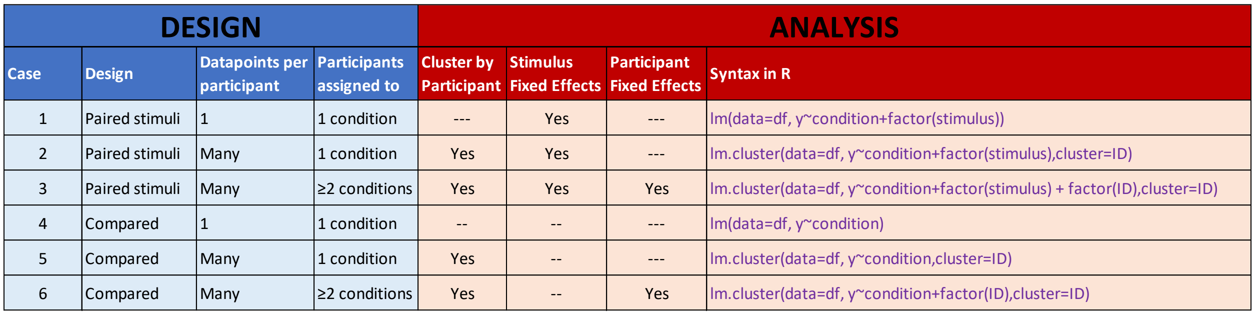
Figure 7. Estimates of overall effect and its confidence interval for the three examples.
R Code to reproduce the figure: https://researchbox.org/2257/10 (use code CXUWHS)

It's interesting to contrast the insights one arrives at when looking at these summary overall results alone, to those one arrives at when inspecting the Stimulus Plots. In Example 1, the regression and mixed-models point to a robust, substantial, and significant effect of intentionality on divulging secrets. Only looking at the Stimulus Plots do we realize a surprising share of stimuli did not work, and only looking at the Stimulus Plots is our attention driven to outlier results that may have been impacted by confounding intentionality with other differences across the two scenarios (e.g., desire to help a person who cuts himself on purpose). In Example 2A, the regression and mixed results both lead to a relatively tight confidence interval for no effect. We would conclude, as the original authors did, that "pose did not impact choice of Black partners" (p.59). Only looking at the Stimulus Plot do we realize that, actually, pose impacted the *majority* of Black partners but that sizeable effects of opposite sign cancel out on average.

Finally looking at Example 3, on the one hand, the Stimulus Plot does not actively add information, for it shows lack of (surprising) variation across stimuli, and thus the overall summary from the regression/mixed models is interpretable as is. But on the other hand, only with the Stimulus Plot do we realize the overall average is an apt summary of the results.

Choosing the statistical analysis

Because regressions with fixed effects or clustered errors are not currently commonly used in psychology, in Figure 8 we provide a brief overview mapping design decisions to the elements that must be included or are beneficial to include in the regression. Specifically, clustered errors prevent inflated false-positive rates due to lack of dependence, while participant fixed effects can increase power, and stimulus fixed effect can also increase power if participants see a subset of all stimuli in a condition. We created this overview to help researchers determine the 'right analysis' for their designs. Which design to choose, however, should be based on substantive considerations related to the particular research question at hand. We note that for binary dependent variables, the mixed model is particularly sensitive to violations of the strong assumptions it is based on (Grilli & Rampichini, 2015; Heagerty & Kurland, 2001), providing a stronger justification to rely on regressions with clustered errors.



DESIGN				ANALYSIS			
Case	Design	Datapoints per participant	Participants assigned to	Cluster by Participant	Stimulus Fixed Effects	Participant Fixed Effects	Syntax in R
1	Paired stimuli	1	1 condition	---	Yes	---	lm(data=df, y~condition+factor(stimulus))
2	Paired stimuli	Many	1 condition	Yes	Yes	---	lm.cluster(data=df, y~condition+factor(stimulus),cluster=ID)
3	Paired stimuli	Many	≥2 conditions	Yes	Yes	Yes	lm.cluster(data=df, y~condition+factor(stimulus) + factor(ID),cluster=ID)
4	Compared	1	1 condition	--	--	---	lm(data=df, y~condition)
5	Compared	Many	1 condition	Yes	--	---	lm.cluster(data=df, y~condition,cluster=ID)
6	Compared	Many	≥2 conditions	Yes	--	Yes	lm.cluster(data=df, y~condition+factor(ID),cluster=ID)

Figure 8. Regression analysis for multi-stimuli studies

For stimuli fixed effect, the same stimulus id must be used for a pair across conditions (e.g., stimulus[1]="face_23", or stimulus[1]="mug"). If authors feel the need to run a logistic regression with clustered errors, they can use: glm.cluster(…, cluster="ID", *family="binomial"*). The function lm.cluster() is included in the 'miceadds' package, it requires indicating the data.frame which is done by including its name as the first argument in the call (see "data" in each row). Other R packages that produce clustered errors include 'lfe', 'jtools', 'estimatr' and 'fixest'.

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 is that when psychologists 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 latent 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. It is very important to stress from the start, however, that Mix-and-Match does not *require* GenAI. Whenever we discuss how researchers can rely on GenAI for a particular step of the process, we also discuss how they may rely on alternative tools if it is more appropriate for the task at hand. Much like using scenario vs lab experiments, monetary vs non-monetary incentives, video vs 'remember a time' inductions of emotions, there are advantages and disadvantages to relying (partially) on GenAI to generate stimulus vs relying on other means. As GenAI becomes mainstream, research practices will adjust and incorporate it, just like they have adjusted and incorporated previous technological developments, without fully eliminating older approaches to designing studies.

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.

(i) Identifying the experimental paradigm. We define the term 'paradigm' as the description of an experimental procedure that constitutes a practical and valid test of a hypothesis of interest, where every specified design element in the paradigm is necessary for the experiment to be a valid and practical test. 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, e.g., that participants shall read texts of a certain length that describe disgusting scenes. 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."

One could add further specificity to this description, e.g., indicating that the experimental paradigm involves reading a passage from the book 'Trainspotting', or evaluating the morality of president Trump kissing his daughter on national TV, but these additional specifications are not justifiable by theoretical concerns (e.g., other inductions of disgusts are equally justifiably ex-ante) nor by practical concerns (it is easy to implement an experiment where different participants read different book segments) thus this description is too specific to meet the definition of 'paradigm'.

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).[14] What we propose, instead of random sampling, is stratified sampling. We propose creating possibly arbitrary categories in the universe of stimuli, strata, that are meaningfully different from each other. Categories should differ on a dimension that corresponds to the instantiation of the latent construct of interest rather than a secondary attribute. For example, creating categories of disgust-inducing passages of text that differ *in the origin* of such disgust: sexual, rot, pest, etc., rather than in the length of the text or some other superficial feature. Mixing involves only alternative operationalizations of the latent independent variable.

As a default, we propose creating 5 strata but if authors have a reason to choose a different number they should. From each stratum, in turn, 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.

[14] A sample of vignettes may not be representative of a population neither in the general statistical sense of the word representative (i.e., used to define a random sample), nor in the sense proposed by Brunswik (1955), where the distribution of stimuli in psychology experiments represents the distribution of stimuli participants might encounter in everyday life.

It is *not* a problem if the strata are not exhaustive (e.g., that the 5 strata that are defined do not capture all 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 producing 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 categories from 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 consists of relying on generative artificial intelligence (GenAI) tools like ChatGPT. GenAI provides an easy, stimulus blind, and documentable approach for implementing the stratification, and sampling, of categorical stimuli. We provide more details about the implementation of GenAI based generation of stimuli because it is more novel an approach than, say, consulting the literature or relying on pilot studies. But we want to stress that we do not think it is intrinsically superior or preferable as an aid for stimulus generation. Like natural intelligence, artificial intelligence can suffer from bias and blind spots that researchers need to keep in mind as they choose stimuli.

To categorize the universe of stimuli, we propose the following "stimulus categorization instruction" to be given to research assistants blind to hypothesis, pilot participants, or as a prompt to a GenAI agent: "*please generate 5 categories of <stimulus universe> that differ in <dimension used to create categories> and provide two specific examples of <stimuli> for each.*" It may help to provide an example of one category and stimulus within it. That second placeholder within that instruction, '*dimension used to create categorie*s', is the more challenging one to specify; experimenters need to consider on what aspect they want the multiple stimuli to vary, striving to generate stimuli that are meaningfully different, hopefully entailing quite different instantiations of the latent variable of interest (rather than the same instantiation differing across stimuli on a superficial attribute). The examples below and in the supplement[15] may be useful for appreciating its role. We posed that "stimulus categorization instruction" as the following ChatGPT prompts for our toy 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).*

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 were: "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 provided involved segments from the books "The Road", "The Sisters Brothers", "The Shining", "Haunter", and "Lolita", respectively.

[15] Supplement available https://researchbox.org/2257/47 (use code: CXUWHS).

Answers obtained for this stimulus categorization instruction, whether given to natural or artificial intelligence agents, include a random component, thus the same instruction may lead to different results over time. Moreover, different researchers may operationalize it differently. This idiosyncratic variability is again fine because the goal is internal validity, not generalizability. We return to this issue in the general discussion when commenting on replicability and reproducibility.

Numerical stimuli. For numerical stimuli, for example monetary outcomes or probabilities, we propose 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, so that every stimulus in one condition is paired with a matched stimulus in another condition, providing multiple mini-replications within a study. Matching can be achieved through three alternative study designs: (i) treated-stimuli design, (ii) paired-stimuli design, and (iii) compared-stimuli design. The first two obtain that ideal of individually matched stimuli.

In treated-stimuli 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. For example, experiments examining how a given item is valued when being bought vs sold, or how a given fake story is treated when it has been previously encountered vs when it is encountered for the first time, are experiments that *treat* stimuli. In treated-stimuli design, the stimuli are naturally paired: treated vs untreated versions of the same stimulus. It's worth noting, however, that *treatments* may be confounded. In the power posing example discussed earlier (Karmali & Kawakami, 2023), for instance, while the same potential partners were in an expansive vs contractive pose, we do not know -because we do not have access to the photos- whether pose-expansiveness is the *only* way in which the photos of the same individual differed, or whether photos of people from different race differed only on the race of the people depicted.

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 "stimulus verification check" we propose later can be used for such purposes.

In paired-stimuli and compared-stimuli 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 answer the following 'confound recognition question': *"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 should 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 considered when matching stimuli.

This confound recognition question *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 recognition question 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 the confound recognition question. Their proposed confounds may instead 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 paired-stimuli design, pairs of stimuli across conditions are formed by matching a target stimulus, say the word "bear", to a matched control stimulus, the words 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. If there is no close enough non-homophonic word neighbor to "bare", then it is not used in the study.

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 stimuli' design, where a set of stimuli in one condition are 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.

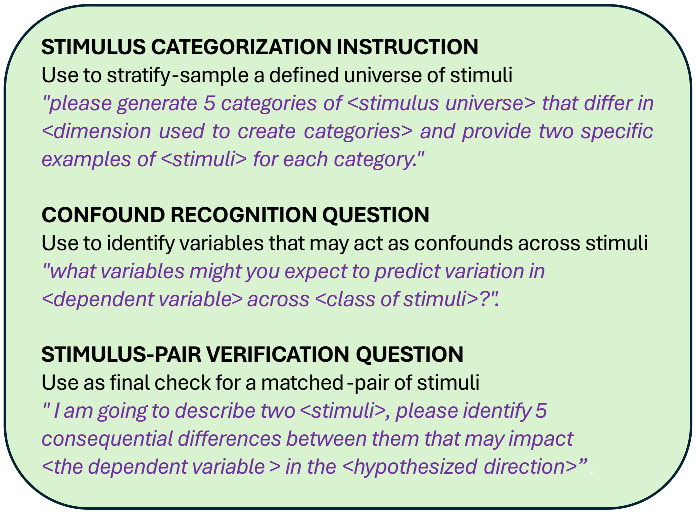
For paired-stimuli designs, we propose a final check to validate pairs, posing the following "stimulus-pair verification question". This question can also be posed to research assistants blind to hypothesis, pilot study participants, or a GenAI agent: "*We are going to describe two <stimuli>, please 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 involving John cutting himself, 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> and

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 (We have included the full text of the answers in Supplement 6). Armed with this feedback it seems straightforward to iterate and modify the scenario to reduce potential confounds (and then submit it again to the stimulus-pair verification question, repeating until potential confounds are all deemed too far-fetched or actually are part of the hypothesized mechanism).



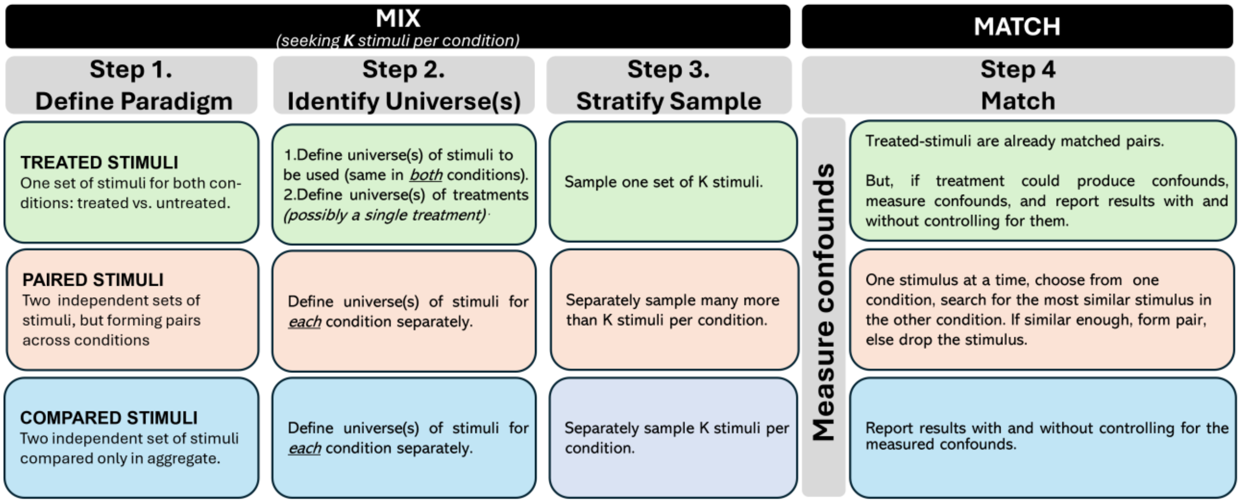
STIMULUS CATEGORIZATION INSTRUCTION
Use to stratify-sample a defined universe of stimuli
"please generate 5 categories of <stimulus universe> that differ in <dimension used to create categories> and provide two specific examples of <stimuli> for each category."

CONFOUND RECOGNITION QUESTION
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>?".

STIMULUS-PAIR VERIFICATION QUESTION
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 9. *Three Key Questions to Pose to Hypothesis-Blind Agents for Mix-and-Match*

Finally, Figure 10 provides a flowchart that summarizes 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.
PAIRED 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 10. Overview of Mix-and-Match

Examples of Mix-and-Match

In Supplements 1-4 (https://researchbox.org/2257/47 | use code: CXUWHS) 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. Here we highlight some design changes that arose from that exercise.

Starting with the study by Salerno and Slepian (2022) on revealing secrets about unintentional vs intentional acts. The mix of (20) scenarios chosen by the authors seemed similarly diverse to the set we produced relying on Mix-and-Match (diverse in domain, in type of immoral act and in reason to act intentionally/unintentionally). The matching within scenario-pair, however, between intentional and unintentional versions of the scenarios, seemed to improve with Mix-and-Match. Specifically, several pairs in the original article often involved substantially different acts, e.g., the aforementioned scenario with John cutting himself to alleviate self-harm vs while chopping vegetables. In contrast, using the Mix-and-Match approach to generate stimuli would have prevented this issue. In addition, with Mix-and-Match, one could produce a larger set of scenario pairs (say 40) and then select the 10 or so that fared best in the 'stimulus-pair verification questions'.

Turning to the study by Karmali and Kawakami (2023) on body posing. It seems to us that a treated-stimulus design that relied on computer generated images that were identical except for skin color would enhance both mixing (obtaining a more diverse set of people) and matching (ensuring only race perception was manipulated), but we explore how Mix-and-Match could be relied upon to modify the original design which relied on real photographs.

In terms of mixing, Mix-and-Match led us to a similar set of poses as the one used by the original authors. In terms of matching, however, it helped us identify room for improvement. The original authors put some effort on matching photographs (see page 53), documenting that the full sets of pictures of Black and White targets had similar *average* age and attractiveness. But there was no matching for *individual* targets between their expansive vs contractive pose photographs (to ensure the expansive and contractive photos differed only in that pose per-se), nor across individuals of different races (the original study had a compared- rather than a paired-stimulus design for race). Following Mix-and-Match one would ideally do a paired-stimulus design, matching each White target with the most similar (on all confounds) Black target, and if this did not prove practical or feasible, one would analyze the data generated by the compared-stimulus design with regressions that did and did not control for the confounding variables.

Finally, turning to the study by Pretus et al. (2023) on factchecking misinformation on Twitter. Implementing Mix-and-Match did not lead us to suggest modifications that would improve on matching, the treated-stimulus design they relied on, where the same piece of false information was shown with or without a fact-check was already matched. However, following the steps of Mix-and-Match did lead us to generate a design with stronger mixing on three counts: (i) the topics about which the fake-news posts were written (ii) the type of misinformation present, and (iii) the type of fact-check. The latter two seemed particularly interesting to us. Our exercise led us to suggest considering, in addition to false-information, misinformation that involved (1) false context, (2) misquotes, (3) manipulated/altered content, and (4) misleading connections. Our exercise led us to suggest considering, in addition to the default Twitter fact-check, (1) in-line annotations, (2) links to fact-checking articles, (3) visual indicators, (4) pop-up warnings and (5) expert commentary.

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 definition of *external* validity, the findings need to be documented… …in the real world.

What about 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.[16] 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.

What about costs? Another concern people may have is that there are experimental paradigms where stimulus sampling could be prohibitively expensive. This is indeed true, especially for field experiments where each stimulus has a large implementation cost. For instance, it may not be possible to attempt 20 or even 5 alternative implementations of a "nudge" in a single field experiment. When practical circumstances prevent using multiple stimuli in a study, one could rely on other (presumably lab) studies, in the same paper, or elsewhere, that rely on stimulus sampling to validate the stimulus used in the field study.

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

[16] To get an intuition for this: imagine two stimuli, one has a very big effect, the other no effect. Using only one of them, blindly, expected power is 52.5%, if you choose the right one you find p<.05 for sure, if you choose the wrong one you only have a 5% chance. But if you use both, with a big enough sample, power is 100%. More generally, if only some stimuli show detectable effects, and ex-ante is hard to tell which (which we intuit is a quite common scenario), using multiple stimuli can dramatically increase power. This intuition we have that anticipating effect size is difficult may seem to be at odds with our concern that experimenters choose confounded stimuli by simulating the experiment in their heads. The premise is that the effect size variation across stimuli with vs without blatant confounds is typically larger than the variation in effect size among ex-ante valid stimuli.

studies, which is valuable for internal and perhaps external validity. Second, running the 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). Fourt, as mentioned above, multiple stimuli 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. See next point.

What's the impact on replicability and reproducibility? Let's begin with replicability, the ease with which researchers can run a new study that seeks to replicate the original one. If the new study is a direct replication, using identical or nearly identical stimuli, then obviously relying on Mix-and-Match is inconsequential. The new study will use whichever stimuli were originally used and run the new study, how the stimuli were originally selected is not relevant for direct replications.

For conceptual replications, which may vary the stimuli used, Mix-and-Match enhances replicability because by articulating the paradigm and the universes of stimuli, readers can more readily extend a particular design to consider new stimuli, and other readers can assess the extent to which the original paradigm has been followed.

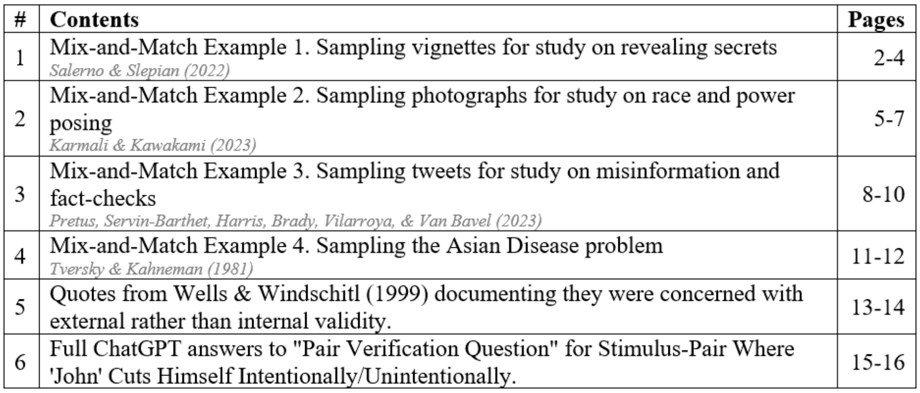
In terms of reproducibility, of repeating the process by which the stimuli were generated, the argument is similar. Mix-and-Match makes this task simpler as well. It is rare nowadays for a paper to provide sufficient details for how stimuli were chosen in order for new researchers to verify that a described process leads to the set of stimuli used. Clearly Mix-and-Match makes this task easier.

Doesn't mediation analysis solve the internal validity problem? 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.

Possible misuses. In this paper we have proposed new tools, and all tools from pencils to rearview mirrors, can be misused. We believe misuse may involve mixing and/or matching over superficial dimensions, so that studies do not actually include truly diverse stimuli nor match stimuli on potential confounds. We don't believe it is possible to fully prevent this, but we hope that the concrete recommendations from Mix-and-Match will reduce unintentional instances of this problem.

In terms of 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 perspective we have given throughout the article will be effective protection against such misuse.

We close with the table of contents to the supplementary materials.



#	Contents	Pages
1	Mix-and-Match Example 1. Sampling vignettes for study on revealing secrets *Salerno & Slepian (2022)*	2-4
2	Mix-and-Match Example 2. Sampling photographs for study on race and power posing *Karmali & Kawakami (2023)*	5-7
3	Mix-and-Match Example 3. Sampling tweets for study on misinformation and fact-checks *Pretus, Servin-Barthet, Harris, Brady, Vilarroya, & Van Bavel (2023)*	8-10
4	Mix-and-Match Example 4. Sampling the Asian Disease problem *Tversky & Kahneman (1981)*	11-12
5	Quotes from Wells & Windschitl (1999) documenting they were concerned with external rather than internal validity.	13-14
6	Full ChatGPT answers to "Pair Verification Question" for Stimulus-Pair Where 'John' Cuts Himself Intentionally/Unintentionally.	15-16

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

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics, 138*(1), 1-35.

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.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3), 255-278.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological methods, 23*(3), 389.

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.

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.

Grilli, L., & Rampichini, C. (2015). Specification of random effects in multilevel models: a review. *Quality & Quantity, 49*, 967-976.

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.

Heagerty, P. J., & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika, 88*(4), 973-985.

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.

Oberauer, K. (2022). The Importance of Random Slopes in Mixed Models for Bayesian Hypothesis Testing. *Psychological science, 33*(4), 648-665.

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.

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208-1214.

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.

Wickens, T. D., & Keppel, G. (1983). On the choice of design and of test statistic in the analysis of experiments with sampled materials. *Journal of verbal learning and verbal behavior*, *22*(3), 296-309.