This version: 2024 08 03
Newest version: https://urisohn.com/47

The Wisdom of Plots: Stimulus Plots Tell You Whether Average Effects Are Interpretable, Evidence From Wisdom of Crowds Experiments

Abstract

We argue that in multi-stimuli experiments, stimulus-level results are more informative than are aggregate results (e.g., than the overall effect of the manipulation computed with a mixed-model). We make this point concrete by revisiting a couple of Psychological Science papers on the "Wisdom of Inner Crowds", where a commentary deemed an original finding spurious after re-analyzing the data with a "maximal" Mixed-Model. We construct Stimulus Plots (which depict stimuli-level results) for the two larger studies in the set. We find that the overall mean computed by the mixed-model is uninterpretable (arguably in one case, for certain in the other). Lastly, also with Stimulus Plots, we explain two violated assumptions on which the "maximal" Mixed-Model relies, which lead to: over-estimating confidence interval width, low power, and a false-positive rate near 0%.

Data and code to reproduce all results are available from
https://researchbox.org/3321 (use code RDCAZB)

Introduction

Experiments with multiple stimuli are typically analyzed focusing on the overall effect between conditions, averaging across stimuli. In this article, we demonstrate that this average effect can be uninformative and misleading. Assessing the interpretability of average effects requires examining stimulus-level results. We illustrate with a recent controversy published in this journal regarding studies on the 'Wisdom of Inner Crowds' (Fiechter, 2024; Van de Calseyde & Efendić, 2022).

The Wisdom of Inner Crowds occurs when the averaging of multiple numerical guesses provided by the same person increases accuracy.[1] For example, if Alex is asked for the year in which the United States declared independence, and Alex first answers "1775", and then "1777", each guess would be off by 1, while the average guess would be off by *less* than 1 (by zero).

Imagine an experiment testing whether that 'wisdom' increases when participants consider, before the second guess, how a person different from them would answer that question. Participants in such an experiment would be randomly assigned to receive an instruction to think about a different other, or not, and would provide estimates for, say, six different numerical questions. The results section describing that experiment may read something like: "comparing the two experimental conditions, we found the predicted effect on accuracy, d = .21, p = .0008". The aspect of that statement that interests us, is that it reports the *average* effect size for the treatment across the six stimuli. That average could be informative and apt, but it could also be uninformative and misleading. We illustrate this in Figure 1 which shows three different scenarios that could be behind the same overall mean (cf. Anscombe, 1973; Gelman, Hullman, & Kennedy, 2023).

[1] Work on the wisdom of the inner crow includes Gaertig and Simmons (2021); (Herzog & Hertwig, 2009, 2014; Vul & Pashler, 2008; Winkler & Clemen, 2004)

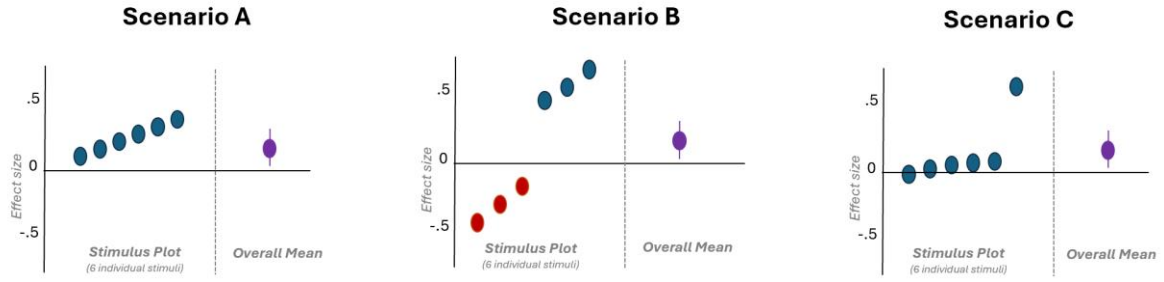

Fig 1. *Stylized illustration of average effects that are, and are not, apt summaries of the data*

Scenario A in Figure 1 is probably the one that comes to mind when reading that the manipulation "improved accuracy, d=.21, p=.0008". In this scenario the average stimulus, the purple circle, provides a good representation of the underlying data and all stimuli exhibit, at least directionally, the summarized effect of greater accuracy with the provided instruction. The true effect sizes for all stimuli may be similar or identical, with random variation in the observed effects around the overall mean.

But Scenarios B & C are plausible alternatives. Here the average effect is not representative; most stimuli exhibit effects quite different from it. In B half the stimuli show an effect in the opposite direction. The statistical summary of the study, that the treatment increases wisdom, is contradicted by half the data. In C only one stimulus shows an effect. What is concerning is that unless the results are reported at the stimulus level, there is no way of knowing which scenario is behind an overall mean.

Over the past 60 years, psychologists have proposed increasingly sophisticated statistical approaches to *account* for variation across stimuli when analyzing the overall effect of a manipulation (i.e., computing the confidence interval or *p*-value of the overall effect). At first, the methodological calls were for running two separate ANOVAs, one for variation across participants, the other across stimuli (Clark, 1973; Coleman, 1964). Then the calls were for relying on mixed-models, which treat the stimuli as drawn at random from a larger set with an assumed distribution (Baayen, Davidson, & Bates, 2008; Judd, Westfall, & Kenny, 2012). Then the calls were for relying specifically on 'maximal' mixed-models (Barr, Levy, Scheepers, & Tily, 2013). In the context of between-subject experiments with multiple stimuli, maximal

mixed-models involve accounting for different stimuli not only having different averages of the dependent variable ("random intercepts"), but also different effects of the manipulation ("random slopes"). There has also been some pushback against relying on maximal models (Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017) and the need for mixed models more generally (McNeish, Stapleton, & Silverman, 2016). Most recently, the calls have been for relying on *Bayesian* maximal mixed-models (Fiechter, 2024; Oberauer, 2022), in part, to address convergence issues with the estimation algorithms. Returning to Figure 1, the focus of this influential and increasingly technical literature, has been on computing 'the correct' confidence interval for those right-most purple circles.

Here we propose redirecting the field's attention to the circles to the left of those purple ones; towards the "Stimulus Plot" (Simonsohn, Montealegre, & Evangelidis, 2024) with stimulus-level results. Concretely, we propose shifting the field's attention toward exploring and understanding qualitative variation across stimuli, and away from merely trying to *account* for it inside a black box, focusing just on the overall mean. If one does not understand why, say, half the stimuli show opposite effects, one doesn't really understand the study results, and one does not really have an interpretable average effect. Statistical precision around an uninterpretable average is not useful, no matter how much mathematical sophistication went into computing it.

We illustrate revisiting a series of studies on the wisdom of inner crowds published in this journal. Specifically, Van de Calseyde and Efendić (2022) proposed, as in our stylized example above, that the wisdom of inner crowds was stronger when people were instructed to answer the way a person very different from them would answer. Fiechter (2024) reanalyzed the same data, using Bayesian maximal mixed models, and concluded that all evidence was spurious. We shall see that Stimulus Plots allow more nuanced takes on these data, moving us towards a more productive and psychologically rich read of the data.

Similar Average Results, Dissimilar Stimulus Plots

Van de Calseyde and Efendić (2022) published results for 5 studies. We focus on the two studies with the most stimuli: Studies 1a and 4, where participants provided two numerical guesses, for each of 10

and 12 numerical questions respectively. The key manipulation was whether participants were asked to think about how a person very different from them would answer prior to providing the second guess.[2] The impact of the manipulation was measured by the difference in accuracy (mean squared error) between a participant's first and average guess for a given question (for a numerical example, see footnote[3]).

Fiechter (2024) notes that while the overall average effects in these studies were significant, p=.02 and p<.001 (see his Table 1), they are actually "spurious" (see his abstract), as these effects go away (have confidence intervals that include 0) when computed with maximal mixed models (the ones that seek to account for each stimulus having a different effect size). Fiechter relied on Bayesian models. In order to obtain p-values, we ran equivalent *frequentist* maximal mixed-models, obtaining p=.12 and p=.255, respectively (in Supplement 1 we show that the Bayesian and frequentist confidence intervals are qualitatively identical). In other words, we reproduce Fiechter's results. One can debate which mixed model is preferrable, maximal vs not, and indeed we do that in the next section, but first one should establish whether the average effects being computed are meaningful. Should one even care whether those overall means are significant?

Figures 2 and 3 help us answer that question. They depict Stimulus Plots for both studies. Focusing on the aggregate results, the right-side of each figure, we see that the mean is positive in Figure 2 and negative in Figure 3. We will discuss that later. More importantly, we see the patterns we just discussed: overall averages are significant with the mixed-model used by Van de Calseyde and Efendić (2022), but not with the maximal mixed-model like the one used by Fiechter (2024). But that's not the only thing we see, and it is not the only thing we should care about.

[2] Study 4 includes another condition where participants were asked to think how someone similar to them would answer. The results for that condition are similar to the condition without thinking about another person.

[3] For example, if the true answer were 10, and the two guesses were 6 and 16, the first guess would be off by 4, while the average guess ((6+16)/2=11) would be off by 1, and so the dependent variable would be $4^2 - 1^2 = 15$. In English, the average guess reduced squared error from the first guess by 15.

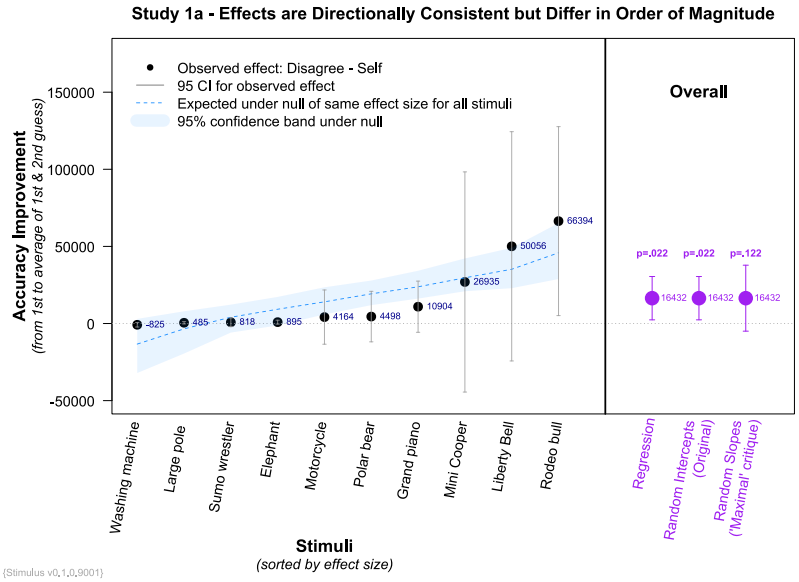

Fig 2. *Stimulus Plot for Study 1a by Van de Calseyde and Efendić (2022)*
The study had 10 numerical questions answered by every participant (N=880) twice. For each participant the squared error of the first and average estimate was computed, and the difference (first answer's MSE – mean answer's MSE) constitutes the dependent variable. The individual stimulus results are obtained with t-tests. The overall average effects are computed with a regression (with errors clustered by participant), and with a mixed-model with and without random slopes. The expected under the null line and confidence band is obtained via rerandomization (shuffling the stimulus ID column).
R Code to reproduce figure: https://researchbox.org/3321/4 (use code RDCAZB).

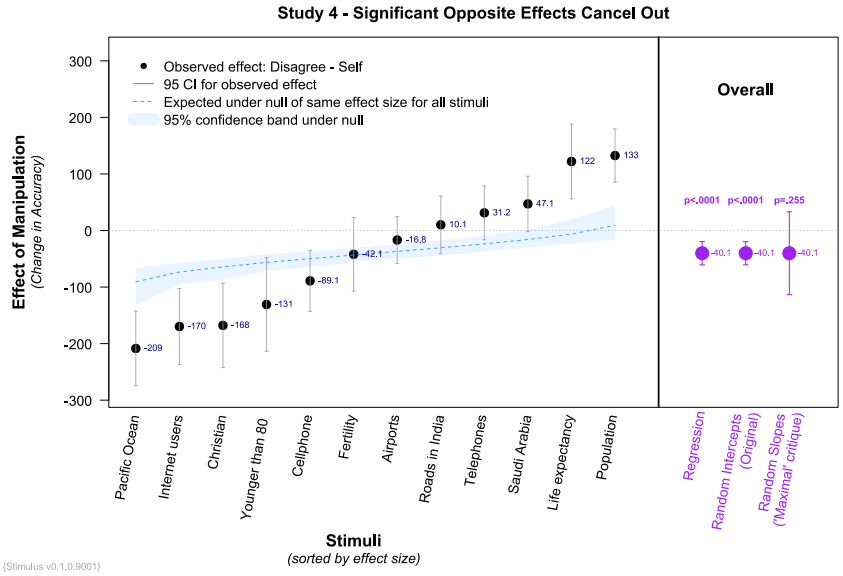

Fig 3. *Stimulus Plot for Study 4 by Van de Calseyde and Efendić (2022)*
The study had 12 numerical questions answered by every participant (N=1836) twice.
For dependent variable calculations see caption for Figure 2.
R Code to reproduce figure: https://researchbox.org/3321/4 (use code RDCAZB).

The individual stimuli results to the left of these overall average are quite informative. Specifically, they suggest that the average effect in Study 1a is of difficult interpretation (for it averages effects of different orders of magnitude), and that the average effect in Study 4 is utterly uninterpretable (for it averages significantly positive and significantly negative effects).

Interpreting the Stimulus Plot for Study 1a: effect sizes of different orders of magnitude

We see in Figure 2 that nine of the ten stimuli show an effect in the same direction. While this is encouraging, the magnitudes are incommensurate. The "Rodeo bull" question has an estimated effect that is 100 times larger than that of each of the smallest four estimated effects. This is concerning both statistically and conceptually.

Statistically, the variability is concerning because the presence of such disparate values may invalidate the assumptions behind the mixed-model calculations. In fact, in Supplement 2, we show that the false-positive rate, which should be 5%, is 0%(!) for the model run by Fiechter; that may sound like a good thing, but it isn't, for it means the test is greatly underpowered and the confidence intervals are wrong.

Conceptually, the variability is concerning because it suggests the psychological effect of interest is being inappropriately operationalized by the chosen dependent variable (change in mean-squared-error). It's *possible* that there is something psychologically meaningful about the Rodeo bull question that makes its effect size 100 times larger, but it seems more likely that the psychological effect is not that different, and that the variability reflects instead idiosyncrasies with the chosen dependent variable (e.g., that people were generally less accurate when answering it, thus showing bigger squared errors and bigger changes in squared errors). Our read of Figure 2, then, is that it would be useful to analyze the data modifying the dependent variable to increase the statistical and psychological validity of the results. One option we considered, depicted in Figure 4, was measuring the effect as a *percentage* change in absolute error.

Reassuringly, the qualitative nature of the results was similar, addressing both our statistical and conceptual concerns. The individual stimulus effects range now between -1.6% and +2.4%, a more commensurate range of values. On the one hand, having survived addressing our concerns, the Stimulus

Plot in Figure 3 helped us be more confident about the overall finding. On the other hand, while one should not expect every stimulus to show a significant effect in the predicted direction, here only 1 stimulus shows an individually significant effect. Moreover, the average effect is small, <1% increase in accuracy, and barely significant (p=.04).

One possible takeaway is that this study design (combination of stimuli and dependent variable) seems to require a larger sample size of participants to be adequately powered. A less optimistic perspective is that, given that the effect is quite small, whether exactly zero or not may be immaterial and not worthy of additional data. Deciding between these perspectives is, of course, a human judgment call rather than a statistical matter.

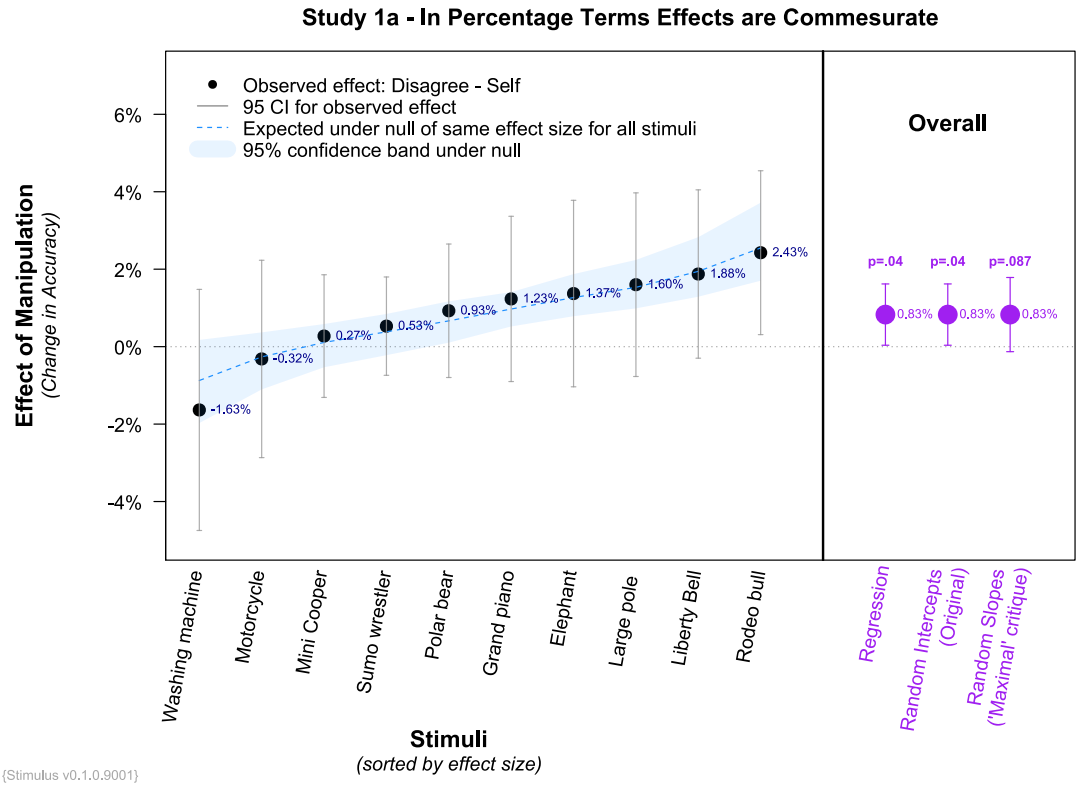

Fig. 4 *Stimulus Plot for Study 1a changing the dependent variable to be percentage change in accuracy*
R Code to reproduce figure: https://researchbox.org/3321/4 (use code RDCAZB).

Interpreting the Stimulus Plot for Study 4: the overall mean is meaningless.

Returning to Figure 3, we begin noting the resemblance to Scenario B in Figure 1: half the stimuli show an effect in one direction and half show an effect in the other. The average effect of such categorically diverse individual stimuli is uninterpretable. It's akin to computing the average effect of drinking water on life expectancy, combining situations where people are thirsty and where people are drowning. Whether the average effect of water consumption is positive or negative in such pool of situations (no pun intended), is meaningless. An overall average of such diverse effects may have appropriate statistical properties, but it has no psychological interpretation.

This heterogeneity in Study 4, it should be noted, was a feature, not a bug. Van de Calseyde and Efendić (2022) purposefully designed the study to have six stimuli they expected to show a positive effect and six which they expected to show a negative effect. In their article they analyzed the sets of stimuli separately (see their Table 3). It was only the critic, Fiechter (2024), who analyzed the data ignoring this prediction of moderation, focusing (exclusively) in the overall average that collapses across stimuli with opposite predicted effects.

Stimulus Plots and The Alleged Necessity of *Maximal* Mixed Models

The core criticism Fiechter (2024) makes to the analyses by Van de Calseyde and Efendić (2022) is that they did not include random slopes in their mixed-models, making their results "anti-conservative" and "spurious" (p.695). He notes that the addition of random slopes to a mixed model "allow[s] analysts to . . . draw more generalizable conclusions at the population level" (p.695). He is not alone in that claim. For example, in an influential article calling for relying on maximal mixed models whenever it is practically possible, Barr et al. (2013) write that models that do not include random slopes "always generalize worse" (abstract).

We agree that if all the model's assumptions were met, a maximal mixed-model would indeed be preferrable. But we believe that in most real-life psychology experiments these assumptions are not met, and in those cases it is unclear whether the maximal-model approach brings about any real benefits. It is

clear, however, that it brings real costs: improper confidence intervals and lower power. It is useful to keep in mind that adding random slopes to a mixed model does not alter the estimated effect of the manipulation. The only change is that the confidence interval (possibly) becomes larger. See for instance, Figures 2, 3 and 4: the model with and without random slopes have the exact same estimate, just different confidence intervals.

We do not think this article about Stimulus Plots and Wisdom of Crowds studies is the right outlet for a detailed evaluation of maximal vs non-maximal mixed models, but with the background we have provided here, we can discuss two assumptions that the maximal mixed model makes, which are often false, and which lead to excessively wide confidence intervals. The assumptions are that (1) stimuli are randomly chosen by experimenters, and (2) effect sizes are distributed symmetrically in the population of stimuli.

Assumption 1. Experimenters choose stimuli at random

Mixed models, both with and without random slopes, assume that stimuli are randomly chosen from a population of all possible stimuli, but only for mixed models with random slopes is the violation of that assumption consequential. This assumption, moreover, is seldom true, for a few reasons. First, it is typically impossible to draw stimuli at random from the population, because such population simply does not exist for most psychology experiments (Simonsohn, 2015). Second, experimenters often choose stimuli purposefully, looking for interesting, unconfounded, convenient, or persuasive stimuli—not at random. The experiment we discussed above, Study 4 by Van de Calseyde and Efendić (2022), exemplifies a third reason why the assumption of random sampling is often false: authors often purposefully rely on *stratified* sampling of stimuli. Specifically, the authors did not choose 12 stimuli at random from the full population of numerical questions (whatever that population would entail), instead, they chose 6 stimuli that they expected to show a positive effect and 6 they expected to show a negative effect. As we shall see, with stratified samples, the confidence interval computed by the maximal mixed-model *overestimates* the true variability of the mean, leading to artificially high *p*-values, and artificially low power.

Let's see why. By assuming the stimuli are randomly drawn from the entire population, the maximal mixed-model will construct the confidence interval by considering the possibility that the experiment could have involved, instead of 6 predicted positive and 6 predicted negative effect, any combination of positive and negative stimuli, say 8 positive and 4 negative, or 3 positive and 9 negative effects. Those *unbalanced* experiments, which would not actually be run, but which are being considered in the calculations, will naturally have associated a wider range of possible average effects, leading to a larger confidence interval for the mean. To be concrete, imagine all stimuli have an effect of size 1 or 2, the negative ones are -1 or -2, and the positive ones +1 or +2. Running 12 stimuli, 6 positive and 6 negative, the range of possible mean effects is between -0.5 and +0.5. But if we allow uneven samples, the range is 4 times larger, ranging between -2 and +2 (see footnote for calculations[4]). The Study 4 confidence interval from the critique by Fiechter (2024), the one that makes $p<.001$ turn into $p=.225$, is over-estimated, among other reasons, because this key assumption of random sampling of stimuli is violated.

Assumption 2. Symmetric distribution

The maximal mixed-model assumes that the universe of possible effects sizes is symmetrical. For instance, that if the average effect is d=.5, then if a stimulus with d=.6 exists, then one should assume that there must be another stimulus out there with d=.4, equidistant from the mean. This is an arbitrary assumption—the central limit theorem applies to means, not to underlying distributions. Why would one expect symmetry in the effect size of a psychological intervention across stimuli? This assumption leads the maximal mixed-model to have properties we believe our readers will find undesirable, even paradoxical.

Let's consider a simulated example with 10 stimuli, see Figure 5. We see that 9 stimuli obtained moderate effect sizes, and one stimulus, 'item 10', has a very large effect, about 5 times that of the others

[4] For an even sample of 6 positive and 6 negative stimuli, the lowest possible mean is obtained if all 6 negative stimuli having an effect of -2, and all 6 positive ones an effect of +1, the resulting average is (-2x6 + 1x6)/12 = -0.5. If in contrast the sample could have all 12 stimuli be negative, each with an effect of -2, the average would be -2.

(this may seem unrealistically extreme, but it isn't, recall that in Study 1a one stimulus had an effect 100 times larger than the four smallest effects).

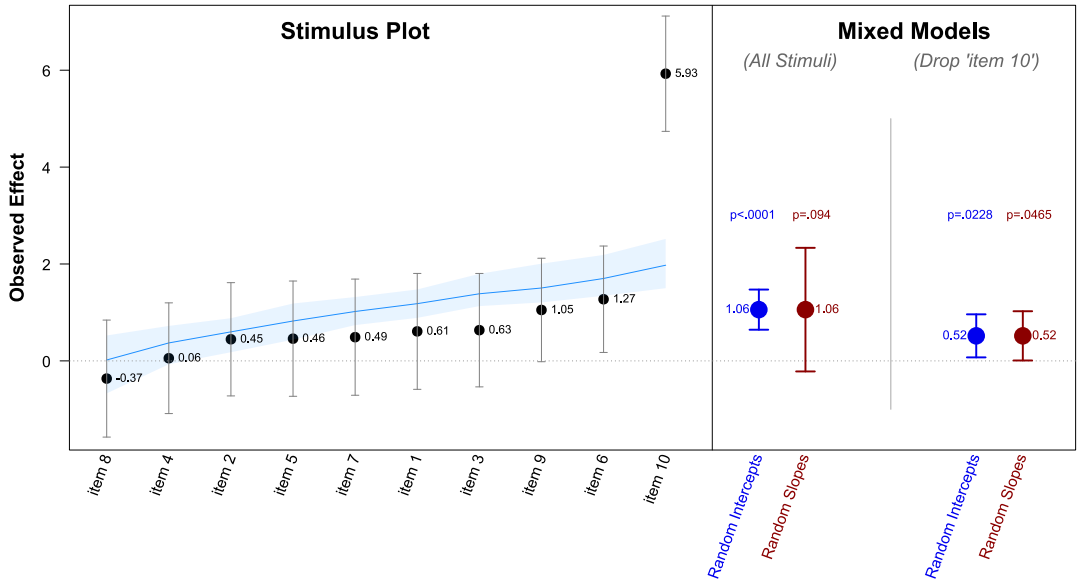

Fig 5. *Mixed-Models with random slopes can paradoxically become less significant after adding a stimulus with very strong effects*
The figure depicts results for a single simulation of 10 studies. Three have a true effect of .35, three of .45, three of .55 and one of 5. Each simulated participant sees all 10 stimuli and has a true random intercept N(0,1). There is also a true random intercept for the stimuli N(0,1). After adding random noise to the dependent variable a mixed model with and without random slopes was estimated on all 10 stimuli, or on 9 excluding 'item 10'.
R Code to reproduce figure: https://researchbox.org/3321/8 (use code RDCAZB).

Let's consider the consequences of including vs excluding that 10th stimulus in the overall mean. If we rely on a mixed model without random slopes (see blue lines), we obtain the pattern we suspect most readers expect. Eliminating data with strong evidence, 'item 10', lowers both the point estimate, from M=1.06 to M=.52 and the statistical significance, from $p<.0001$ to $p=.028$. But the mixed model with random slopes (see red lines) shows a different pattern. Eliminating 'item 10' leads higher significance, going from $p=.094$ to $p=.0465$. The intuition for this paradoxical pattern is that when the random slopes

model sees 'item 10', an effect far above the mean, it imagines other stimuli similarly far from the mean *in the opposite direction*. As a result, it becomes less certain that the average effect is actually positive.[5]

Discussion

Over the past 60 years methodologists have been proposing increasingly sophisticated models to compute confidence intervals around overall effects in experiments with multiple stimuli. We believe those models, and their ongoing debates, have distracted psychologists away from focusing on what matters more: the information provided by individual stimuli. We personally believe maximal mixed-models are not generally justified and cause more harm than they are worth. However, we hope that even readers who disagree with us and remain enamored with the elegance of maximal mixed models and their promise of 'generalizability' will, in the future, accompany their analyses with Stimulus Plots, discussing the extent to which the patterns of results across stimuli are consistent with one another, with the overall mean, and with the underlying hypothesis of interest.

In this article, we have shown how Stimulus Plots can inform the meaningfulness of overall average effects. Stimulus Plots are also useful for identifying unexpected confounds and moderators, by identifying unexpected variation across stimuli. For example, that only a particular kind of stimulus shows the hypothesized effect. For more on this, see the article introducing Stimulus Plots (Simonsohn et al., 2024).

[5] We selected this simulation because it makes a particularly compelling paradox, but to guard against having chosen an extreme random draw, we repeated the exercise 100 times. In 65% of cases the *p*-value from the maximal model was lower when the larger stimulus was dropped, and in 20% of cases the overall effect went from $p>.05$ to $p<.01$. For the mixed-model without random slopes, in contrast, this happened 0 and 1 times, respectively.
See https://researchbox.org/3321/8 (use code RDCAZB).

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*(1), 17-21.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59*(4), 390-412.

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.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

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.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports, 14*(1), 219-226.

Fiechter, J. L. (2024). Drawing Generalizable Conclusions From Multilevel Models: Commentary on. *Psychological science, 35*(6), 694-699.

Gaertig, C., & Simmons, J. P. (2021). The psychology of second guesses: Implications for the wisdom of the inner crowd. *Management Science, 67*(9), 5921-5942.

Gelman, A., Hullman, J., & Kennedy, L. (2023). Causal quartets: Different ways to attain the same average treatment effect. *The American Statistician*, 1-6.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological science, 20*(2), 231-237.

Herzog, S. M., & Hertwig, R. (2014). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 218.

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.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of memory and language, 94*, 305-315.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2016). On the Unnecessary Ubiquity of Hierarchical Linear Modeling. *Psychological methods*.

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

Simonsohn, U. (2015). The Effect Size Does not Exist. Retrieved from http://datacolada.org/33

Simonsohn, U., Montealegre, A., & Evangelidis, I. (2024). *Stimulus Sampling Reimagined: Designing Experiments with Mix-and-Match, Analyzing Results with Stimulus Plots [preprint] (http://dx.doi.org/10.2139/ssrn.4716832)*.

Van de Calseyde, P. P., & Efendić, E. (2022). Taking a disagreeing perspective improves the accuracy of people's quantitative estimates. *Psychological science, 33*(6), 971-983.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological science, 19*(7), 645-647.

Winkler, R. L., & Clemen, R. T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis, 1*(3), 167-176.