Revision of XGE-2017-072 as invited by the action editor, Nelson Cowan

Causal Processes in Psychology are Heterogeneous

Niall Bolger, Katherine S. Zee, Maya Rossignac-Milon

Columbia University

Ran R. Hassin

Hebrew University of Jerusalem

Author Note:

Correspondence regarding this article should be sent to Niall Bolger, Department of Psychology, Columbia University, New York, NY 10027.

Word Count: 10,668

Commented [NB1]:

Abstract

All experimenters know that human and animal subjects do not respond uniformly to experimental treatments. Yet theories and findings in experimental psychology either ignore this causal effect heterogeneity or treat it as uninteresting error. This is the case even when data are available to examine effect heterogeneity directly, in within-subjects designs where experimental effects can be examined subject by subject. Our goal in this paper is to convince experimentalists to change their assumptions about causal processes in psychology from a one-size-fits-all view to one that allows for subject-specific experimental effects. Using data from four repeated-measures experiments, we show that causal effect heterogeneity can be modeled readily, that its discovery presents exciting opportunities for theory and methods, and that allowing for it in study designs is good research practice. We conclude by urging experimentalists—at all stages of their research process—to work from the assumption that causal effects are heterogeneous.

Commented [KZ2]: We also use the term experimentalists in this abstract—make consistent?

Keywords: causal processes, theory development, heterogeneity, repeated measures, mixed models

Causal Processes in Psychology are Heterogeneous

All organisms within a population show intrinsic heterogeneity. They vary from one another to some degree in structure and function. Darwin, in *The Origin of Species* (1865), considered this heterogeneity an essential basis for natural selection and evolutionary change. That heterogeneity exists in phenotypic features such as size, shape, color, and symmetry, is therefore neither surprising nor unnatural. It is true for microorganisms, plants, and animals; and humans are no exception.

In much experimental work, however, heterogeneous *responses to treatments* are regarded as random error—background noise that obscures the signal of an experimental effect. This view can be traced to terminology present in classic works on experimental design such as Fisher's *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935). The deeper roots of this terminology lie in the research topics that engaged statistical pioneers such as Gauss and Laplace in the 17th century (Hald, 1998; Stigler, 1986). Because their work dealt with problems of astronomical measurement (such as determining the exact position of stars through repeated measurements), it was natural to label variation in measurements as *error*, or chance deviations from a true value. To this day, the view that average values are truth and variations around the average are error is deeply ingrained in experimental disciplines in the biological and social sciences.

It is typical practice to focus on averages within experimental conditions rather than variability (except indirectly through significance tests). This approach can make sense if that variability can be attributed to nuisance factors such as measurement error, irregularities in application of the experimental procedure (treatment error), or fleeting states of participants (e.g., momentary lapses in attention). If, however, the variability includes *true differences between participants in responses to experimental conditions*, then there can be lost opportunities for understanding the phenomenon, and perhaps most importantly, for constructing adequate theories. There may also be undesirable consequences for research practice. Failures to incorporate causal effect heterogeneity can result in false conclusions about the efficacy of experimental treatments.

What is Causal Effect Heterogeneity?

These differences in experimental effects form the central concept of this paper, that is, *causal effect heterogeneity*: variation across experimental units (e.g., people) in a population in the size and/or direction of a cause-effect link. Although it is often neglected in theory and research in experimental psychology, it has become a fundamental concept in modern treatments of causality beginning with Rubin (1974) and increasingly in social research (see, e.g., Angrist, 2004; Brand & Thomas, 2013; Gelman & Hill, 2007; Imai & Ratkovic, 2013; Molenaar, 2004; Western, 1998; Xie, 2013). In this literature, causes are defined as within-unit comparisons across experimental conditions and assumed to vary in size across experimental units in a population.

Commented [RH3]: Not the clearest sentence I have read…

In typical between-subjects experimental designs, where one observation is obtained on each subject, causal effect heterogeneity, if present, cannot be distinguished from non-causal sources such as measurement error and treatment error. However, within-subjects repeated-measures designs, in which causal inference involves comparing subjects to themselves in other experimental conditions, offer unique opportunities to examine this heterogeneity directly. Most repeated-measures experiments conducted in cognitive psychology, social psychology, and other areas provide sufficient data to allow for individual-specific experimental effects. Conceptualizing an experimental effect as variable rather than constant opens the door to theory development.

Causal Effect Heterogeneity in Experimental Psychology

Beginning with Estes (1956), one can find influential papers that draw attention to causal effect heterogeneity and how conventional models based on group averages provide an inadequate account of psychological processes (see also, Estes & Maddox, 2005; Lee & Webb, 2005; Whitsett & Shoda, 2014).

~~Relatedly, more recent examinations suggest that group summary statistics can fail to capture important information about individuals (Fisher, Medaglia, & Jeronimus, 2018).~~ Despite ~~this~~ thisese ~~early and more recent work~~~~perspectives~~prior work—and with notable exceptions discussed below—experimental psychology has generally neglected this topic. A prevailing assumption appears to be that causal effect heterogeneity is either absent or, if present, is irrelevant to theories of psychological processes.

A contributing factor to this neglect has been the traditional reliance of experimental psychologists on repeated-measures ANOVA to analyze within-subjects effects. Repeated-measures ANOVA, at least as it is implemented in popular software, makes it difficult, if not impossible, to estimate causal effect heterogeneity[1]. Linear mixed models are needed to do so (McCulloch, Searle & Neuhaus, 2008). Mixed models have their roots in biostatistics (e.g., Henderson, 1953). Broadly applicable and flexible versions of mixed models emerged in the 1970's and 80's in the form of new computer algorithms and software (Fitzmaurice & Molenberghs, 2009), and their use has grown exponentially since then. Mixed models are also known as multilevel, mixed-effects, or hierarchical regression models (Raudenbush & Bryk, 2002; Gelman & Hill, 2007; Maxwell, Delaney & Kelley, 2018; Snijders & Bosker, 2011).

Commented [MR4]: CONVERTED ALL FOOTNOTES TO ENDNOTES PER NIALL'S COMMENT BELOW

Commented [NB5]: The more I think about it, the more I'm inclined to make the footnotes into endnotes. As footnotes they distract the reader--or this reader at least!

Papers advocating the use of mixed models, whether based on Frequentist (Baayen, Davidson & Bates, 2008; ~~Hamaker & Wichers, 2017;~~ Hoffman & Rovine, 2007; Locker, Hoffman & Bovaird, 2007) or Bayesian (Lee & Webb, 2005; Rouder & Lu, 2005) principles, have pointed the way forward for experimentalists. Two areas in experimental psychology that routinely use these models are experimental linguistics (for which the Baayen et al., 2008, has been a major influence) and the field known as cognitive modeling that has its roots in mathematical psychology (Busemeyer & Diederich, 2010; Lee & Wagenmakers, 2014).

Commented [NB6]: The Hamaker reference wasn't to experimental data.

Commented [KZ7]: Hamaker does talk about the need to use MLM with experimental data, though. See page

Commented [KZ8]: Could cite Hamaker 2017 here?

Beyond these exceptions, though, only a small fraction of current work in experimental psychology takes causal effect heterogeneity into account adequately. Of the papers appearing in issues 1-6 of *JEP: General* in 2017 that used repeated-measures designs (50 total), nearly two-thirds (62%) used repeated measures ANOVA to model their data. Of the 38% of papers using mixed models, only 9 indicated whether individual-specific effects (random slopes) were estimated[2]. Importantly, only 2 papers reported on the estimates themselves.

Aims

This paper has three broad aims. The first is to convince experimentalists in psychology to change their metatheory of causal processes from a one-size-fits-all view to one that allows for subject-level heterogeneity. We will show how causal effect heterogeneity, when present, has important and exciting implications for theory development and empirical testing.

The second aim is to provide the field with an accessible guide on how to estimate, display, and draw theoretical implications from causal effect heterogeneity. Although the guide is intended for future work, it also can be used for exploration of the many existing repeated-measures datasets that have the ability to shed light on causal heterogeneity were they modeled appropriately.

A third aim is to show that attention to causal effect heterogeneity will lead to better research practice. In doing so, we will show how experimentalists can use mixed-models for repeated-measures data in new and fruitful ways. In addition to presenting opportunities for theory, adequately modeling heterogeneity can help protect researchers against underpowered studies, illusory findings, and replication failures. A salient example of the problem of illusory findings is a retracted *Psychological Science* paper by Fisher and colleagues (2015), whose primary experimental effect was undermined ~~disappeared~~ once causal heterogeneity was modeled properly.

Because our primary concern is with theory formulation and testing, we advocate models that are adequate for this task, that are readily available, and that are easy to use. Some causal processes in experimental work will no doubt require the more sophisticated tools that are now becoming available, but in our view, the bulk of the benefits can be obtained using simpler approaches.

In sum, the goal of this paper is to demonstrate the utility for theory, methods and practice of incorporating causal effect heterogeneity in repeated-measures experimentation. To do so, we will address four specific questions:

1. How can causal effect heterogeneity be estimated? (Study 1)
2. When is the heterogeneity sufficiently large to have implications for theory or sufficiently small to be ignorable? (Studies 2 & 3)
3. Can theoretically relevant variables explain the observed heterogeneity, and what are the implications of the heterogeneity for further experimental investigations? (Study 1, revisited)
4. Is the heterogeneity ephemeral or enduring? What does this imply for theory and for future experiments? (Study 4)

Study 1: Estimating Causal Effect Heterogeneity

As noted, most datasets from repeated-measures experiments are suitable for examining causal effect heterogeneity. We begin by illustrating how heterogeneity can be estimated for a specific research question. For this and for subsequent questions, in the Supplemental Materials we provide data, analysis code in R and SPSS, and outputs for researchers to explore and to serve as templates to follow in their own work.

Effect of stimulus valence on reaction time for self-descriptive traits

Our first example dataset comes from a conceptual replication of Study 1 of Scholer, Ozaki, and Higgins (2014), in which participants were presented with positively and negatively valenced trait words and asked to indicate whether each of the words was self-descriptive. Response time for each word was measured. A straightforward prediction is that participants will be faster to endorse positive self-descriptions, given that people are motivated to maintain a positive self-view (Leary & Baumeister, 2000; Yamaguchi et al., 2007). We chose this hypothesis because we wanted to examine possible heterogeneity in a robust and well-documented effect.

Participants

Seventy-five students from Columbia University participated for 1 course credit or $5. The sample size was nearly triple that used in the original study on which it was based. Thirteen were excluded for failing an attention check, leaving a sample of 62 participants.

Procedure

Procedures were approved by the Columbia University IRB; procedures for other studies were approved by the IRBs of the institutions where those data were collected. After giving consent, participants were led to individual cubicles to begin the experimental task, which was administered on a computer with PsychoPy (Peirce, 2007). Participants completed the Regulatory Focus Questionnaire (Higgins et al., 2001), additional individual difference measures that we will not discuss further, and general demographic questions [3]. Next, participants completed a computerized task measuring the trait valence effect. Finally, participants were debriefed, compensated, and thanked.

Commented [WU9]: I worry that the footnotes, especially the long ones will be unnecessarily distracting to readers. How about we use endnotes instead?

Measure of trait valence effect

Each trial began with a fixation point that appeared for 1 second, followed by a trait word. Twenty words were of positive valence (e.g., “talented”, “disciplined”), and twenty were of negative valence (e.g., "boring", "impulsive"). The participants' task was to indicate whether they possessed the trait or not, as quickly as possible, by pressing a designated key on the keyboard. The trait word disappeared when the response was made, and 2 seconds later the next trial began. The first 6 trials served as a practice phase, followed by 40 experimental trials. Each trait appeared once, in random order for each participant. The computer recorded the response latency (i.e., the time elapsed between the appearance of the word and the key-press) as well as the yes/no response (i.e., whether the word was endorsed as self-relevant or not). See Section 1 of Supplemental Material for details of a pilot study conducted to determine the trait words.

Mixed model analysis and visualization

As is common with reaction time (RT) data, we used the natural log transformation to remove skewness (although using raw RT scores did not change the results). Only trials containing words endorsed as self-relevant were included in the analyses, in accordance with procedures used by Scholer et al. (2014). There were 3 participants who did not endorse any negative words. Thus, analyses drew on data from 59 participants. On average, participants endorsed 22 words as self-relevant, 62% of which were positively valenced~~; however~~. T~~, t~~here was a substantial range, however, with participants endorsing as few as 13 words as self-relevant and as many as 28 words. To examine our hypothesis of valence effects on logRT, we used a statistical model where, for each subject, valence was the single experimental ~~condition~~ manipulation and reaction time in log-milliseconds was the outcome. This model allowed us to examine whether people, on average, respond faster when endorsing positive vs. negative self-relevant traits, while also allowing us to examine the variability in this effect[4]. Syntax and output for this analysis are shown in Supplement 2.

Statistical model

Our analysis approach is similar to a standard repeated-measures ANOVA with a single within-subjects factor with repeated trials within factor levels (see Maxwell, et al., 2018, for a description of classic repeated-measures ANOVA).[5] Rather than using repeated-measures ANOVA, however, we use a mixed or multilevel modeling approach. As noted earlier, mixed-models can reveal the existence and size of causal effect heterogeneity.

Our model specifies that in the population, a typical subject's reaction time to endorse self-relevant traits is a function of trait valence, but it also allows subjects to vary in the strength and even the ~~sign~~ direction of the causal effect. Because it is a generalization of repeated-measures ANOVA, it provides the usual test statistics reported in a repeated-measures analysis (on condition main effects and contrasts), and in the case of no missing data on Y, it gives identical results (Maxwell et al., 2018).

We describe the model in terms of the distribution of Y rather than the more conventional linear equation for Y. The two descriptions are interchangeable, but the distribution form makes it easier to see the assumptions of the model (see Stroup, 2012). There are three distributions. The first specifies the distribution of trial-level logRT, and it has parameters for subject-level means and valence effects. The second and third specify between-subject distributions for these subject-specific mean and valence effect parameters.

$$logRT_{ij} \sim N(\mu_j + \beta_j X_{ij},\ \sigma_\varepsilon) \tag{1}$$

In (1) above, the logRT observed for subject *j* for the stimulus in trial *i* is drawn from a normally distributed population with a subject-specific mean function and a subject-general standard deviation. The subject-specific mean function is composed of a parameter μ_j, the subject's model-predicted grand-mean logRT (the subject's overall level or *random intercept,* in the language of mixed models), and a parameter β_j, the subject's causal effect of valence (the *random slope*). Specifically, β_j is subject *j*'s difference in logRT between positively and negatively valenced stimuli, where stimulus valence X_{ij} is coded -0.5 if the stimulus is negative and 0.5 if the stimulus is positive. The common standard deviation, σ_ε, refers to the (residual) variation in logRT scores within each valence condition within each subject.

The distributions of the subject-specific parameters are presented in (2) and (3).

$$\mu_j \sim N(\mu,\ \sigma_\mu) \quad (2)$$

$$\beta_j \sim N(\beta,\ \sigma_\beta) \quad (3)$$

In (2), the subject-specific levels (random intercepts), μ_j, are specified to be normally distributed around mean μ that represents the population average logRT and standard deviation σ_μ that represents heterogeneity (i.e., between-subject variability) in levels. In the language of mixed models, the mean μ is called a *fixed effect* and σ_μ is the standard deviation of a *random effect* of subjects (McCulloch, et al., 2008).

In (3), the subject-specific causal effects of valence (random slopes) are specified to be normally distributed around mean β that represents the population average causal effect and standard deviation σ_β that represents heterogeneity in causal effects. Note that the model also allowed for a correlation between the heterogeneous levels (intercepts) (μ_js) and heterogeneous causal effects (slopes) (β_js); this was omitted to simplify the exposition, but it is included in the analysis.

The software syntax required to estimate the model in R and SPSS is provided in the Supplemental Material. Although in the body of the paper we present conventional Frequentist parameter estimates, in the Supplemental Material we present equivalent Bayesian versions based on noninformative priors (for R software only).

Results

Table 1[6] summarizes the key estimates of interest, namely, of the population parameters from Equations 2 and 3. Estimates are indicated by the use of a caret or hat symbol (^) over each parameter. We will focus on the heterogeneity of the causal effect of valence (bolded). The typical subject ($\hat{\beta}$) is -0.16 logRT units (approximately 150 ms) faster at responding to positively-valenced words than to negatively--valenced words~~, *b* = -0.16, *t*(51) = -7.29, *p* < .001~~. The 95% confidence interval (CI$_{95}$) ranges from -0.21 to -0.12 logRT, which is evidence of a robust effect in the population.~~The 95% Confidence Interval (CI$_{95}$) for this estimate is [-0.21, -0.12], indicating that the causal effect for the typical person in the population is roughly between -0.2 and -0.1 logRT units.~~ [7]

Commented [KZ10]: There are just different algorithms. They are iterative, and the results will converge on something similar. Diffs come in at the 3rd decimal place or so. But all are doing maximum likelihood.

The heterogeneity parameter, the standard deviation of the subject-level causal effects, is 0.13, which is almost as large as the average causal effect.

Formatted: Space Before: 12 pt

We saw that Equation 3 specified that each subject in the population had his or her own causal effect of valence, β_j~~beta-j~~ and that these effects were normally distributed in the population with fixed-effect mean ~~Given that~~ β ~~= -.16~~ and random-effect standard deviation σ_β. With sample estimates of these of these parameters~~= 0.13~~ in hand, we are in a position to calculate a 95% ~~Population~~ Heterogeneity Interval (~~P~~HI$_{95}$) for the valence effect, which captures the range of experimental effects that can be expected in the population. ~~-$\beta_j$beta-j:~~ It ranges from ~~are parameter estimates of a normally distributed causal effect variable, we can infer that 95% of the population includes causal effects as negative as -0.16 – 1.96 (0.13) =~~ -0.41 to ~~units and as positive as -0.16 + 1.96 (0.13) =~~ 0.09 (~~units~~-0.16 +/– 1.96*0.13; see Table 1) and shows that a once-size-fit-all view of the valence causal effect is a mistaken one. At the lower extreme, the model predicts that there are subjects whose causal effects are more than twice that of the average subject, whereas at the higher extreme there are those with no effect or a small reversal.

Field Code Changed

Field Code Changed

The ~~P~~HI$_{95}$ for the distribution of the valence effect β_j~~beta-j~~ should not be~~is interval should not be~~ confused with ~~a~~ the CI$_{95}$ for the mean of that distribution, β~~beta-hatj~~. The ~~P~~HI$_{95}$ concerns predicted *subject-to-subject* variability in the causal effect in the population (using the current sample estimates).~~,~~ ~~and it is of particular interest in this paper.~~ The CI$_{95}$ for the valence effect β, by contrast, concerns ~~*(hypothetical)*~~ *sample-to-sample* variability in estimates of the average causal effect in the population. Both intervals are important, but they answer different questions. The question of how much subjects differ in a causal effect in the population is quite distinct from the question of how precisely estimated the causal effect for the average person is.~~, and it--or functions of it--are usually the focus of experimentalists. It is well known, for example, that when the lower bound of CI$_{95}$ for an experimental effect does not include zero the effect has a *p*-value less than 0.05.~~ ~~95% CI. The 95% CI is an estimate of *sample-to-sample* variability in estimates of the *average* subject's causal effect in a population (were one to conduct the study many times). The 95% heterogeneity interval is an estimate of *person-to-person* variability in causal effects in a population (using the estimates from the current study).~~

Formatted: Space Before: 12 pt

Field Code Changed

Field Code Changed

Formatted: Font: Italic

Field Code Changed

Formatted: Font: Italic

Formatted: Font: Italic

Table 1.

~~Summarized~~ Multilevel Model Output for Trait Valence Effect~~Dataset~~, in logRT units (Study 1)

Effect	~~Population~~ Parameter Estimates			95% ~~Population~~ Heterogeneity Interval	
	Mean	SD		2.5%	97.5%
Intercept (or Avg. Level): $\hat{\mu}_j$ ~~($\hat{\mu}_j$)~~ CI$_{95}$	6.87 [6.82, 6.91]	0.16 [0.13, 0.20]		6.54	7.19
Slope (or Causal Effect): $\hat{\beta}_j$ CI$_{95}$	**-0.16** [-0.21, -0.12]	**0.13** [0.08, 0.17]		**-0.41**	**0.09**

Mixed models of repeated-measure data usually provide two types of output. The first are estimates of *population* parameters for fixed (constant) and random (varying) effects. These are the fundamental mixed-model results, and we have just discussed these above. ~~presented these in numeric form.~~ The second type is *predictions* of effects for each *subject* in the sample. In Frequentist mixed models (the approach taken here) these are called *Empirical-Bayes (EB) Predictions* or sometimes *Best Linear Unbiased Predictions (BLUPs)* (McCulloch et al., 2008; Rabe-Hesketh & Skrondal, 2004). In Bayesian mixed models, they are called hierarchically shrunken estimates (see, e.g., Kruschke, 2015).

Visual displays of both the population estimates and sample predictions can greatly increase understanding of causal effect heterogeneity. Figure 1 shows a strip chart where~~.~~ t~~T~~he x-axis displays a range of values for the causal effect of valence ~~slope~~. First, the population results: The vertical red line (labeled A) in the center is the fixed effect of valence, the model's best guess as to where the population average effect lies. We saw in Table 1 that this value is -0.16. The area between the vertical red lines (labeled C) to the left (-0.41) and right (0.09) shows the 95% P~~p~~opulation ~~causal effect~~ H~~h~~eterogeneity I~~i~~nterval (~~P~~HI$_{95}$) for the causal effect already seen in Table 1 ~~(95% interval)~~. Next, the sample predictions: The blue dots represent each subject in the sample, and the blue dashed lines are the 2.5

Formatted: Line spacing: Double

Formatted Table

Commented [KZ11]: Intercept vs. means

Commented [MR12]: Going to keep this mean, so that it's clear that there is a mean intercept and a mean slope.

Field Code Changed

Formatted: Right, Space After: 0 pt, Line spacing: single

Field Code Changed

Formatted: Font: Italic

and 97.5 percentiles for the causal effects the sample (B). Whichever interval we choose to focus on we can see substantial heterogeneity in the experimental effect.

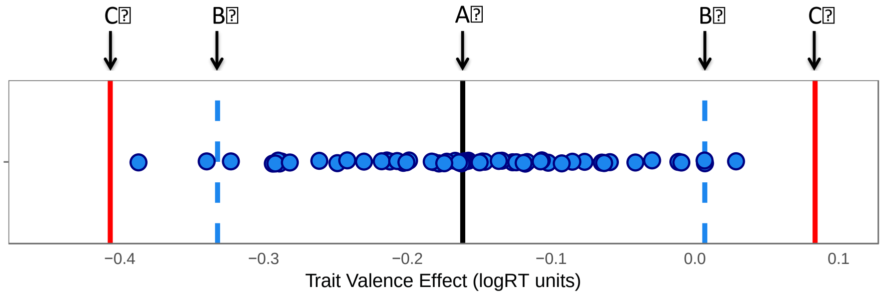


Figure 1. Strip-plot model estimates of between-subject heterogeneity in ~~displaying model estimates of~~ the valence causal effect ~~of valence for each person in the sample~~. A shows the estimate for the population mean; B shows the upper and lower bound containing 95% of causal effects~~of~~ for individuals in the sample~~the sample estimates~~; and C shows the upper and lower bound containing 95% of the range of causal effects in the population (the 95% Population Heterogeneity Interval, HI_{95}, for the causal effect) ~~estimates~~.

A second useful visualization involves overlaying the subject-level predictions on subjects' observed data. Figure 2 is a panel plot showing several subjects' raw data for logRT as a function of valence, with the model-predicted values for each subject. Each fitted line corresponds to a valence-effect data point in the strip chart above. The panels are ordered by the size of the model-predicted valence effect. We display five subjects: the two subjects with the steepest negative slopes, the two subjects with the flattest slopes, and the subject at the median. Note that (a) the model's predictions generally correspond to each subject's raw data and (b) subjects are markedly different from one another. The subject on the far left shows a predicted causal effect that is approximately 2.4 times larger than the subject in middle panel (~~≈~~ -0.39/-0.16). Also, subjects' slopes range from large and negative to slightly positive.[8]~~.~~

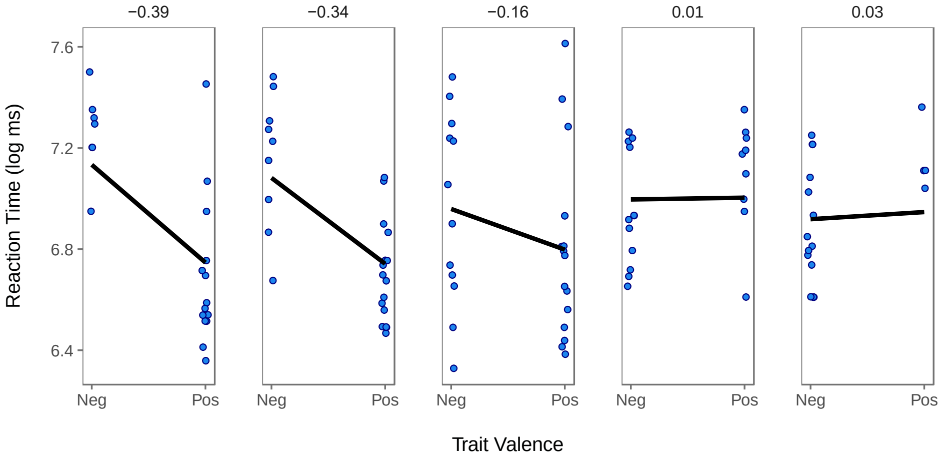


Figure 2. *Panel-plots showing several subjects' raw data for log RT as a function of valence, together with the model-predicted values for these subjects. Values above each plot reflect the size of the valence effect for that subject.*

Advantages of Mixed Modeling Approach

We have just seen how a mixed-modeling approach allows us to estimate and display causal effect heterogeneity. One might ask, though, whether the model-~~-~~predicted effects are any better than simply calculating the valence effect for each ~~participant~~ subject separately (and ~~e.g.,~~ running a one-sample *t*-test for each ~~participant~~subject). This alternative ~~e problem is that the latter~~ approach, however, can give the mistaken impression of heterogeneity even when none exists in the population. Consider the case of a single subject in our study. ~~We obtained RTs repeatedly from the subject for the negative and positive conditions.~~ The mean difference between the subject's responses across conditions is, in itself, an unbiased estimate of the subject's causal effect. Its true value is uncertain to some extent, however, because we used only a limited number of trials within each condition. That uncertainty is indexed by the standard error of the subject's mean difference, and one can think of it as a form of measurement error.

Formatted: Font: Italic

Now consider viewing the effect for a sample~~et~~ of subjects, each of whose experimental effect is uncertain. Just as one would see with a set of error-prone measurements, the observed variation will be the sum of ~~overstate~~ the true variation and the error variation, and will always show and upward bias. In our example, the subject-by-subject valence effect heterogeneity must be adjusted downward ("shrunken") in order for it to be a valid estimate of true population heterogeneity. Mixed models provide a way of accomplishing this, and the adjustments needed for the current study are shown in Figure 3. The top row of Figure 3 shows individual-specific observed differences in logRT as a function of valence, ~~such~~ as would be used in paired t-tests~~obtained by running a t-test for each person in the sample~~. The bottom row shows the subject-specific shrunken estimates from the mixed model. The more uncertain a subject's raw mean difference, the more it is shrunken toward the estimated population mean. These were described above as Empirical-Bayes estimates, or hierarchically shrunken estimates (for further detail, see Raudenbush & Bryk, 2002; Gelman & Hill, 2007; Maxwell, Delaney & Kelley, 2018; Snijders & Bosker, 2011).

Commented [KZ13]: We say one-sample t-test above. I know that these end up being the same for RM studies, but maybe we should stick with one term for consistency?

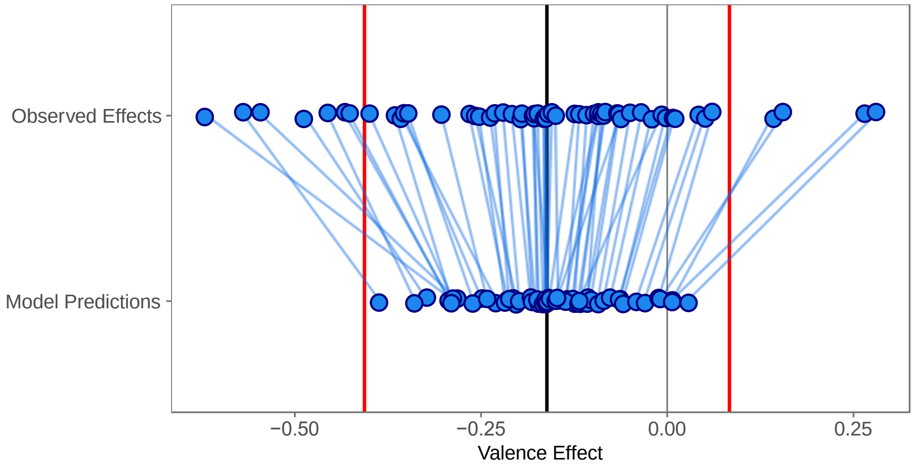


Figure 3. Comparison of observed differences in each participant's valence effect (top-row) and model estimates for each participant's valence effect (bottom row). The solid black line indicates the model predicted average effect. The thin grey line indicates zero point, where a subject is equally fast to ~~0 line (indicating equally fast responses to~~ endorse positive and negative words. ~~) is represented by the thin gray line.~~ The red solid lines show the model's

estimates for th~~[9]~~e 95% heterogeneity ~~population~~ interval in the population.[10]~~.~~ ~~Note that one additional participant could not be included in this visualization because this person's observed difference could not be computed as only one negative trait was endorsed.~~

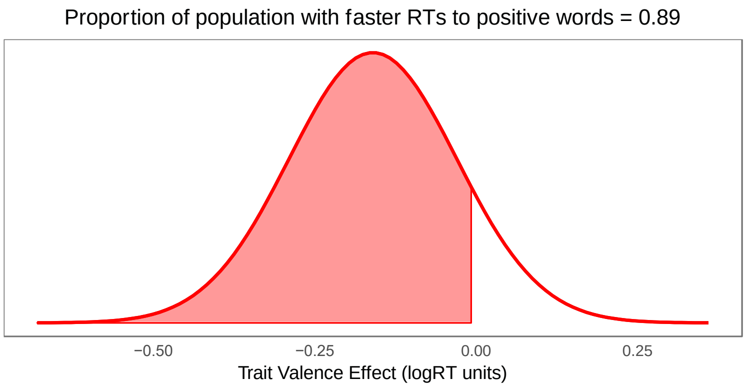


Figure 4. Complete distribution ~~Distribution~~ of subject-specific trait valence effects in the population (based on sample estimates*)~~ for trait valence effect~~. Eleven percent~~The mixed model results suggest that 11%~~ of the population are predicted to show ~~can be expected to show~~ reversals in the valence effect (faster response times to negative words).*

However, even if the sample estimates are shrunken such that reversals are weak or do not occur in the *sample*, the model can indicate whether reversals are likely in the *population.* A useful way of visualizing the latter is to display the population heterogeneity ~~normal~~ distribution implied by the model's estimates of the population mean ~~causal effect~~ (-0.16 units) and SD (0.13 units). As show in Figure 4, the model predicts that ~~shows that distribution for the trait valence example. It~~ predicts that 11% of the population can be expected to show reversals.

Armed with knowledge about the existence and magnitude of heterogeneity, we are now in a better position to communicate our findings. An example of how one might communicate both the average causal effect and the heterogeneity in that effect in a write-up is presented in Supplement 10. In addition, this ~~observed~~ heterogeneity calls for a theoretical account of its existence and magnitude.

Commented [KZ14]: Removed the word "observed" since we used "observed" in Figure 3 to mean something different

What can explain why some participants respond faster to positive traits while others show no difference or even the reverse pattern? Later in the paper we will consider a motivational explanation of the heterogeneity. We will examine whether subject differences in promotion focus, a relatively stable individual tendency to eagerly pursue ideals and aspirations (Higgins, 1998), can explain some of the between-person heterogeneity we found in the trait valence causal effect. ~~Promotion focus has been linked to an emphasis on positives ("gains") and faster response times (Higgins, 1998; Förster, Higgins, & Bianco, 2003), hence its potential theoretical relevance to the heterogeneity we found.~~ However, r~~R~~egardless of whether an investigator can account for it or not, the heterogeneity we observed is fundamental to understanding these experimental results.

Studies 2 & 3: Is Causal Effect Heterogeneity Noteworthy or Ignorable?

We have presented results in which the extent of causal effect heterogeneity was considerable enough to undermine the idea of a common, uniform causal process even though we could be confident that the effect existed for the average subject in the population. Not all experimental phenomena, however, can be expected to show heterogeneity. In this section, we provide two examples, one in which the heterogeneity is noteworthy, and one in which it is not. We also provide guidelines for determining whether the degree of heterogeneity is sufficient to qualify conclusions of repeated-measures experiments.

Noteworthy: Face-Orientation Effects

Study 2 used data from a study by Sklar and colleagues (2017) investigating non-conscious processing speed, specifically the effects of spatial orientation on how quickly participants responded to faces presented using continuous flash suppression (for more, see Sklar et al., 2017). During the study, participants completed trials in which a face appeared on the screen in one of three orientations: Upright, 90-Degrees, and Upside-Down. Participants indicated the orientation of the face, and reaction times were measured. For simplicity, we will focus on the Upright vs. Upside-Down conditions only (.5 = Upright, -.5 = Upside-Down). Our analyses drew on data from 21 participants. As this study and the remaining studies involve secondary analyses of existing datasets, sample sizes were not determined with the present research question in mind. On average, participants completed 121 trails (range = 118-126), yielding 2544 observations total. Trials were roughly equally distributed across the two conditions for each participant. Data are again analyzed in logRT units. R and SPSS syntax and output are available in Supplement 6.

Table 2.

Summarized Multilevel Model Output for Face Orientation Dataset

Effect	Parameter~~opulation~~ Estimates			95% ~~Population~~ Heterogeneity Interval	
	Mean	SD		2.5%	97.5%
Intercept (or Avg. Level) ~~$\hat{\mu}_j$~~ CI$_{95}$	5.11 [4.98, 5.24]	0.28 [0.21, 0.39]		4.56	5.66
Slope (or Causal Effect) ~~$\hat{\beta}_j$~~ ~~**Orientation** ($\hat{\beta}_j$)~~ CI$_{95}$	-0.20 [-0.26, -0.14]	0.11 [0.06, 0.16]		-0.42	0.02

Formatted Table

Field Code Changed

Field Code Changed

Formatted: Right, Space After: 0 pt, Line spacing: single

Formatted: Right, Level 4, Space After: 0 pt, Line spacing: single, Keep with next, Keep lines together

As summarized in Table 2, the average person is -0.20 logRT units faster at responding to an upright versus an upside-down face (CI$_{95}$: [-0.26, -0.14]), with a heterogeneity estimate of 0.11 SD units. This heterogeneity estimate is just over half the size of the fixed effect estimate. We regard this as substantial: These estimates imply that the ~~P~~HI$_{95}$, the 95% Population Heterogeneity Interval for the ~~of the population have~~ causal effects ~~that~~, ranges from -0.42 ~~(-0.20 - 1.96*0.11)~~ to 0.02 ~~(-0.20 + 1.96*0.11)~~ logRT units. A person at the lower bound ~~2.5th percentile~~ shows an effect of face orientation ~~of -0.42,~~ twice as large as that of ~~the effect for~~ the average person, whereas a person at the lower bound ~~97.5th percentile of the distribution~~ shows essentially no ~~causal face orientation~~ effect ~~of face orientation~~. The model's predictions for the actual participants in the sample, as shown in the strip-plot (Figure 5) and panel plots (Figure 6), mirror these population predictions.

Another way to assess the importance of the heterogeneity effect is to compare statistical indicators of model fit for a model with a random slope for face orientation and one without. The results from the comparison suggests the addition of the random effect substantially improves the fit of the model to the data ($\chi^2(2) = 27.9$, $p < .001$). Practically speaking, this test enables us to conclude that a model allowing for heterogeneity in ~~random~~ intercepts and slopes fits the data significantly better than a model allowing for ~~random~~ heterogeneous intercepts only.

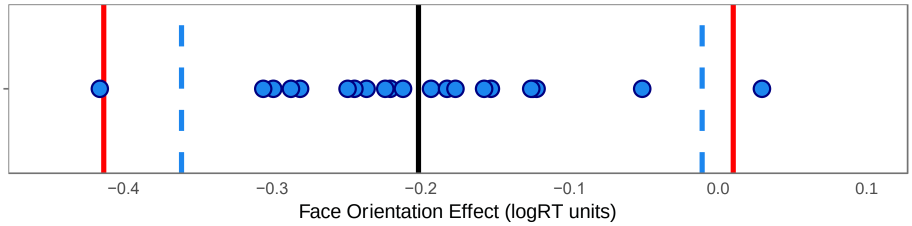


Figure 5. *Strip-plot displaying model estimates of the causal effect of face orientation for each person in the sample. The black line is the average (fixed) effect, the blue dashed lines show the 95% sample interval, and the red solid lines show the 95% population heterogeneity interval.*

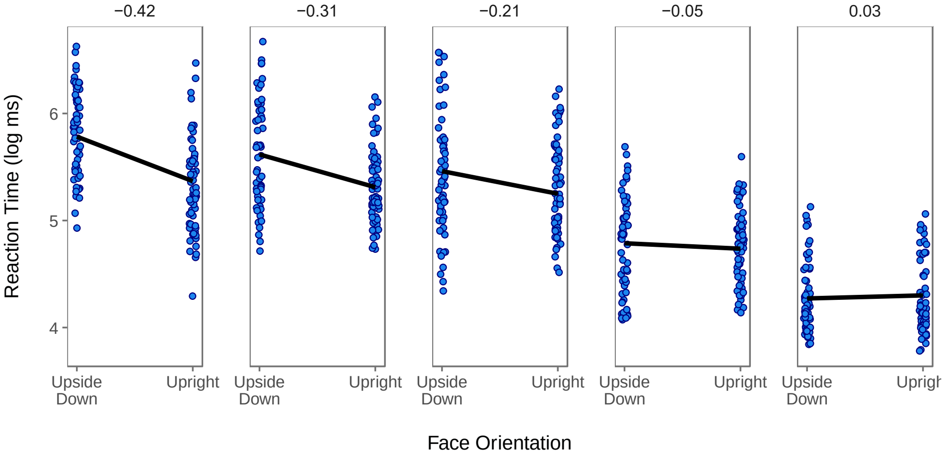


Figure 6. *Panel ~~-~~plots showing several subjects' raw data for logRT ~~reaction time~~ as a function of face orientation, together with the model-predicted slopes for each subject.*

We have now seen evidence of heterogeneity in a second example dataset. The results suggest that focusing exclusively on the mean causal effect, and ignoring the ~~of the~~ causal heterogeneity distribution, will result in an inaccurate picture of the phenomenon. Further empirical work and theorizing is needed to understand why people differ substantially in the face orientation effect. Furthermore, we now know that future studies of the face orientation effect will may need to recruit larger samples, as larger samples are required to conduct adequately powered studies when effects are heterogeneous (see Bolger & Laurenceau, 2013; Snijders & Bosker, 2011).

Ignorable: Math Priming Effects

Although we believe that causal effect heterogeneity is widespread in psychological processes, we acknowledge that in particular instances or in certain areas of research, it may be sufficiently small to be ignored. The next repeated-measures experiment, published as Experiment 6 in a paper by Sklar and colleagues (2012), is such an instance. ~~-~~Heterogeneity was not a focus of the experiment, and the analyses were conducted using repeated-measures ANOVA on aggregated data. Here, our goal is to present a simplified analysis of some of their data using a mixed-model approach to examine heterogeneity in the math priming effect. Thus, this is not a direct reproduction of the analyses and results discussed in the original paper. ~~Here we present a simplified analysis of some of their data using a mixed-model approach.~~

Our Study 3 dataset consisted of 17 participants, who each completed up to 74 trials (range = 67-74 trials, as trials with no response were omitted prior to analysis). A total of 1214 observations were available for analysis, which represent an average of 71 trials/participant. The study examined participants' RTs to pronouncing simple numbers depending on whether subjects were subliminally

Commented [RH15]: This cannot be Study 1. Maybe Experiment 6?

Also, there were a few different conditions – with different presenation times, which do you use?

And lastly, the original paper showed effects only for a subset of the stimuli, that is – for subtraction. I'm assuming that this is what you use?

primed with equations that yield this number (“congruent”) or not (“incongruent”)[11]. The original results showed a substantial congruency effect, indicating that simple subtraction operations are processed and solved non-consciously. For consistency with other studies in this paper, we report analyses using logRTs. The dataset along with R and SPSS syntax are available in the Supplemental Materials; excerpted portions of syntax and output are in Supplement 7.

Table 3.

Summarized Multilevel Model Output.

Effect	Parameter ~~Population~~ Estimates			95% ~~Population~~ Heterogeneity Interval	
	Mean	SD		2.5%	97.5%
Intercept (or Avg. Level) CI$_{95}$	6.48 [6.40, 6.56]	0.17 [0.12, 0.16]		6.15	6.81
Slope (or Causal Effect) $\hat{\beta}_j$ ~~$(\hat{\beta}_j)$~~ CI$_{95}$	-0.022 [-0.040, -0.005]	0.0004 [0, 0.0229]		-0.023	-0.021

As summarized in Table 3, the effect of congruence on logRT is -0.023 units, CI [-0.040, -0.005]. This effect shows essentially no causal effect heterogeneity: The ~~(heterogeneity~~ SD estimate of~~=~~ 0.0004 ~~SD~~ units is less than 2% of the mean value~~)~~. Consistent with this estimate, the ~~We again display predictions and plots to highlight how minuscule the estimated heterogeneity is.~~

~~Based on these results, the P~~HI$_{95}$~~;~~ ~~the population 95% heterogeneity interval~~ for ~~implied population range of effects for~~ the congruence effect is extremely narrow, ~~:~~ from -0.024 to -0.022. The predictions for the sample, displayed in Figures 7 and 8, are in a similar range. ~~These indicate that the heterogeneity effect was less than 2% of the fixed effect value. The strip and panel-plots of these predictions show essentially no variation in slopes.~~ Using the model comparison approach described on

Formatted Table

Field Code Changed

Formatted: Right, Space After: 0 pt, Line spacing: single page 14, the parameter had a negligible ~~-~~contribution to model fit, $\chi^2(2) = .002$, $p = .99$. In this case, we can confidently conclude that the priming effect is~~, in effect,~~ essentially the same across subjects.[12]

Formatted: Do not check spelling or grammar

While one could argue that a mixed-model approach in this case adds nothing beyond what could be found using a repeated-measures ANOVA, note that we now have evidence for the *absence* of a heterogeneity effect. This knowledge will have important implications for power calculations for future studies using the same manipulation and replication attempts by other laboratories (see Kenny & Judd, 2017). One should bear in mind, of course, that

t~~T~~hese results~~, of course,~~ are population-~~-~~specific. Studies of a different population might show substantial causal effect heterogeneity. ~~Results obtained from an effect-homogeneous population can fail to replicate in effect-heterogeneous population.~~

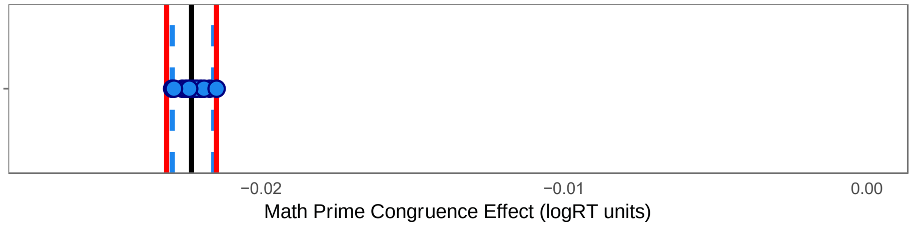


Figure 7. Strip-plot displaying model estimates of the causal effect of prime congruency for each person in the sample.

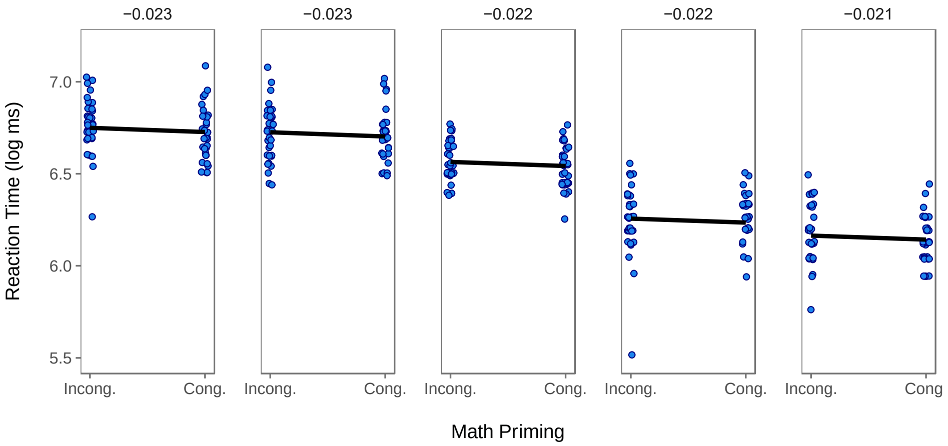


Figure 8. *Panel-plots showing several subjects' raw data for log reaction time as a function of prime congruency, together with the model-predicted values for each subject.*

How to Decide if Causal Effect Heterogeneity Matters

In these examples, we used three criteria to decide whether the causal heterogeneity in an experimental effect was sufficiently large to warrant attention. The first was the *uncertainty interval*, i.e., whether the 95% confidence interval for the heterogeneity parameter included (or was very close to) zero. The face orientation confidence interval suggested 0 heterogeneity was unlikely, whereas the math priming confidence interval suggested 0 heterogeneity was plausible. The second criterion was the *comparative model fit*, i.e., whether the model fit was improved by allowing for heterogeneity vs. not. We saw clear evidence that it was for the face orientation data, but not for the math priming data. The third was the *relative size of the heterogeneity effect*~~, and especially its size~~ in relation to the fixed effect (the effect for the average subject). For the face orientation data, its relative size was 0.50; for the math priming data, it was approximately 0.02~~less than 0.10~~.

Formatted: Font: Not Bold, Italic

Formatted: Font: Italic

We suggest that as a rule of thumb, causal effect heterogeneity is noteworthy if its SD is 0.25 or greater of the average (fixed) effect. Such heterogeneity implies that the ~~P~~HI$_{95}$ includes ~~95% of the population have~~ effect values that lie between 0.5 and 1.5 times the effect for the average person. Thus an individual at the 2.5th percentile of the distribution has an effect size that is half as great as that of the average person, and an individual at the 97.5th percentile has an effect size that is 50% greater than that of the average person. Note that these calculations assume, as we have in (3) above, that the population of causal effects is normally distributed. Other distributions are also possible (e.g., Rouder & Haaf, 2018).

Using this 0.25 threshold, we conclude that the level of heterogeneity in the face orientation data is noteworthy. The random effect of orientation, with an SD of 0.5 of the fixed effect, implies that the HI$_{95}$ ~~95% of the population~~ ranges from ~~has values that lie between~~ 0 ~~and~~ to 2 times the fixed effect. In contrast, for the math priming data, the SD of 2% of the ~~the random effect was 0.0004/ 0.023 = 0.02 of the~~ fixed effect implies that~~, yielding a~~ the HI$_{95}$ ranges from ~~95% population interval between~~ 0.96 ~~and~~ to 1.04 times the fixed effect. Clearly, b~~B~~ased on our threshold~~,~~ this level of heterogeneity is ignorable.

Although we find this relative size criterion to be a useful heuristic, there may be cases in which, based on the goals of the research, researchers may decide to apply stricter or more liberal cutoffs. There are also other approaches that can be used to assess whether heterogeneity is noteworthy (e.g., using Bayes factors; see Rouder, Morey, Speckman & Province, 2012, for an exposition of Bayes Factors in linear and mixed models).

Study 1, Revisited: Explaining Causal Effect Heterogeneity

As we have argued, discovering causal effect heterogeneity, even without knowing its sources, can be a contribution to understanding a phenomenon. Features such as its relative size, whether some subjects showed reversals of sign, and whether the population studied was demographically or culturally homogeneous can have important implications for next steps in a research program. However, if it is the case that researchers included theoretically relevant background measures (e.g., individual differences, demographics, etc.) in a given experiment, the mixed-model analysis above can be expanded to include these measures as explanatory variables.

This section presents an example that draws on existing theoretical knowledge to elucidate the origins of the heterogeneity demonstrated in our first experimental example, in which we found both a robust causal effect of trait valence for the average person (a fixed effect in the language of mixed models) as well as substantial heterogeneity of this effect (a random effect). This kind of heterogeneity can be thought of as a “stand-in” for theoretically relevant explanatory variables. What theories might help us account for why some people respond much faster when endorsing positive (versus negative) traits and why others respond equally quickly regardless of valence?

Drawing on Regulatory Focus Theory (Higgins, 1998), we test the prediction that a chronic (stable) promotion-focused motivational orientation, which involves eagerly pursuing ideals and aspirations, will predict faster endorsement of positively valenced traits. The purpose of this demonstration, as in the demonstration of the overall valence effect, is not to reveal new insights about Regulatory Focus Theory (it is already known, for example, that promotion focus is associated with faster RTs; Förster, et al., 2003). Rather, the purpose of the example is to show of how a theory-derived variable can be used to help explain existing causal effect heterogeneity.

If we consider a generic between-subjects predictor *Z* (e.g., promotion), that is a linear predictor of heterogeneity in grand mean μ_j and the causal effect heterogeneity effect β_j, then distributions (2) and (3) become:

$$\mu_j \sim N(\gamma_0 + \gamma_1 Z_j,\ \ \sigma_\mu) \tag{4}$$

$$\beta_j \sim N(\delta_0 + \delta_1 Z_j,\ \ \sigma_\beta) \tag{5}$$

We will focus on β_j, the causal effect heterogeneity outcome. If *Z* is mean-centered, then δ_0 is the causal effect for the average person (an intercept term), and δ_1 is the effect of *Z* on the heterogeneity (a slope term). The coefficient δ_1 captures the extent to which the heterogeneity effect differs as *Z* differs by one unit. With *Z* taken into account, the standard deviation σ_β is now no longer the total variation but rather the residual variation in heterogeneity. It can be interpreted as how much heterogeneity remains unexplained. To investigate the potential explanatory role of promotion, we estimate the same model as in ~~section~~ Study 1, but now we add promotion focus (mean-centered) as a between-subjects predictor of the heterogeneity, accomplished by allowing promotion to interact with valence[13]. For R and SPSS code and output, see Supplement 8.

The mixed-model results indicate that those with higher promotion scores show a greater tendency to be faster to endorse positive vs. negative traits: = -0.13 logRT units, *t*(60) = -2.89, *p* = .005, CI$_{95}$ [-.22, -.04]. Thus the -0.16 logRT speed advantage of the typical subject is increased to -0.16 – 0.13 = -0.29 logRT for those one unit above the mean on promotion.

To what extent does promotion explain the heterogeneity ~~(i.e., variance)~~ in the valence effect? To answer this question, we must first compute the total ~~valence~~ heterogeneity variance ~~of the valence effect~~ implied by our model with promotion focus. This is akin to calculating the total variance in a regression or ANOVA model using the following formula (see Kutner, Nachtsheim, Neter & Lee, 2005):

$$V(\beta_j) = \delta_1^2 V(Z_j) + \sigma_\beta^2 \tag{6}$$

where $V(\beta_j)$ is the total heterogeneity variance, δ_1^2 is the square of the regression coefficient linking the covariate *Z* to the total heterogeneity (in our example, promotion focus orientation), and σ_β^2, the residual variance in heterogeneity after taking promotion focus orientation into account.

Unlike in linear models such as regression and ANOVA, variance explained in mixed models does not necessarily increase with the addition of predictor variables. Once more variables have been introduced, the model can take this new information into account and provide a revised estimate of variance explained. This is why it is necessary to compute the implied total heterogeneity from a model after including a relevant predictor.

We know from the model output that the heterogeneity of the valence slope in the model including promotion is .013. Using (6) above, we can calculate that the implied total heterogeneity is .018. From there, we can compute the proportion of heterogeneity explained by promotion as: 1 – ($\hat{\sigma}^2_\beta$ / $\hat{V}(\beta_j)$ (i.e., 1 minus the total heterogeneity divided by the residual heterogeneity. In this case, 1 – (.013/.017) tells us that promotion focus accounts for 23% of the between-person heterogeneity in the causal effect.

In Figure 9, we provide a visualization of this effect. The top panel shows the residual heterogeneity (both population estimates and sample predictions) for the model when we include promotion as a predictor. The bottom panel shows the implied total heterogeneity from that model. As expected, the implied total heterogeneity is clearly larger than the residual heterogeneity. In Figure 10, we show how participants' scores on promotion focus predict the implied random effects.

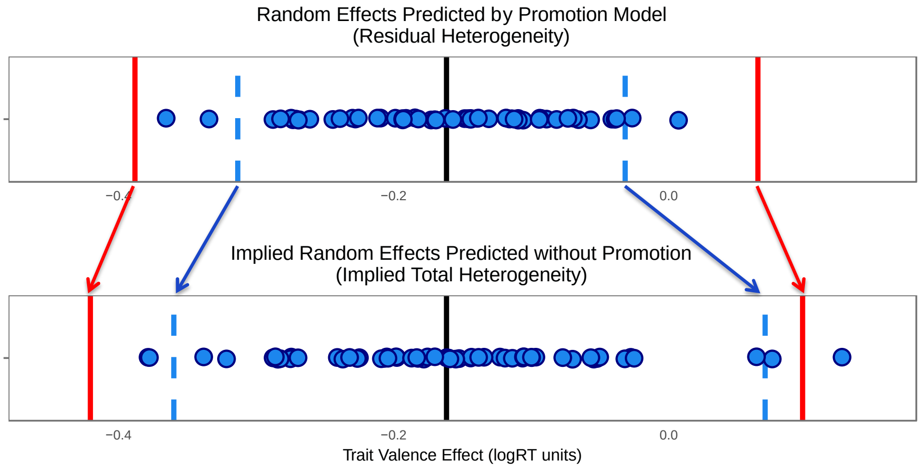


Figure 9. *Strip-plot displaying model estimates of the causal effect (slope) of trait valence for each person in the sample with promotion focus (top panel) and the implied total heterogeneity without promotion (bottom panel). Solid lines show the model-estimated average effect, blue dashed lines show the predicted 95% interval of the sample, and the red solid lines show the predicted 95% interval of the population. As the plots show, the implied total heterogeneity is larger than the residual heterogeneity.*

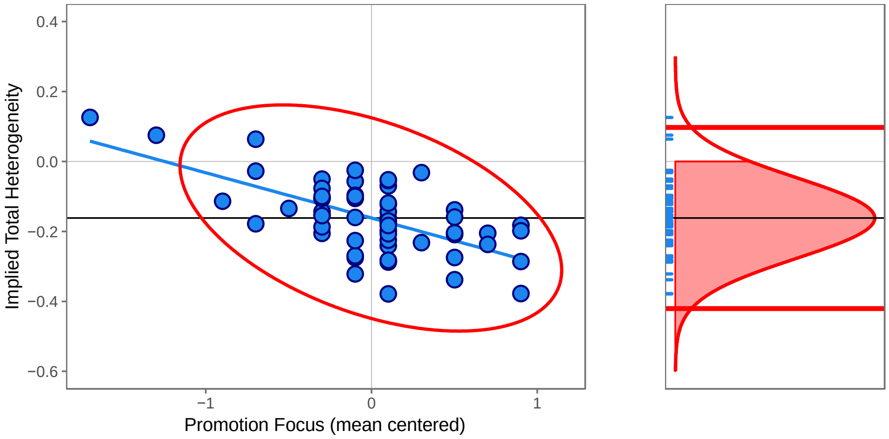


Figure 10. *Scatter plot showing the relationship between promotion focus and the implied random effects predicted by the model, with 95% population ellipse, is displayed in the left panel. A distribution of the implied population effects is displayed in the right panel, along with the mean (black line) and 95% population limits (horizontal red lines). The blue dots in the right panel are the implied random effects for the sample.*

Study 4: Is Causal Effect Heterogeneity Ephemeral or Enduring?

Causal effect heterogeneity can be a function of physical and psychological states that subjects bring to the experimental situation, and that can even endure over the course of the experiment~~, ~~. Such states (e.g., being hungry) ~~-~~might be unl~~but that are not l~~ikely to recur w~~h~~ere the subjects to be brought back for a second experimental session. ~~(e.g., being hungry).~~ But as we have just shown, it can also be at least partially attributable to more stable aspects of participants, such as their motivational orientation. If the causal heterogeneity is due in part to temporally stable characteristics, it follows that the heterogeneity itself should show some temporal stability. In this section, we will investigate this idea in a novel methodological way by examining the temporal stability of causal heterogeneity over the course of a week. This will represent a methodological and modeling innovation that can have important implications for theory development and research practice.

To do this, in Study 4, we involve data from the Scholer and colleagues (2014) paper, in which a sample of Japanese participants completed the trait valence task described in Study 1. However, these researchers administered the trait valence task *on two separate occasions,* one week apart. At Times 1 and 2, participants' reaction times to endorse 40 positive and negative traits as self-relevant were measured[14]. Our examination drew on a sample of 21 participants. The average participant endorsed 20 traits as self-relevant at each time point (Time 1 range = 12-26; Time 2 range = 11-26). A total of 850 observations were used for our analysis.

In their paper, Scholer and colleagues used a repeated-measures ANOVA and found main effects of valence on RT at each time point. The focus of our analysis, however, will be a question not addressed by Scholer and colleagues: the temporal stability of heterogeneity in the valence effect. Thus, we expanded our modeling approach to simultaneously estimate causal effect heterogeneity at Times 1 and 2 and their correlation. The R code and output of the analysis are provided in Supplement 9[15].

Table 4 summarizes estimates of the valence effects and their heterogeneity at each time point. In this table, the reader will note that the average causal effect of the valence manipulation was -0.14 logRT at ~~time~~ Time 1 and -0.19 logRT at ~~time~~ Time 2; the causal effect heterogeneity was 0.19 SD logRT units at ~~time~~ Time 1 and 0.27 logRT units at ~~time~~ Time 2.

Although these changes in level and heterogeneity of the valence effect are worthy of scrutiny, our focus here is on *temporal stability*. Are those participants showing relatively large valence effects at ~~time~~ Time 1 the same people showing relatively large effects at ~~time~~ Time 2?

The answer to this question is yes, and the extent of this stability, displayed in Figure 8, is striking. There is a very close correspondence between a subject's relative positions at each time point. The data points are the predictions for each ~~participant~~ subject and the ellipse is the population 95% confidence ellipse. The correlation between the causal effect heterogeneity distributions at Times 1 and 2 is 0.95[16].

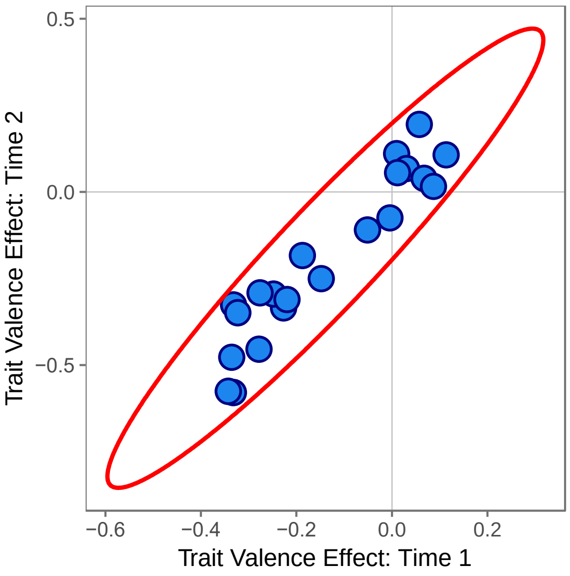
Thus, heterogeneity in this context seems to be attributable to more enduring tendencies, such as promotion focus or other variables, and not to temporary states of participants that endure only over the course of a single experimental session. In other words, this result demonstrates for the trait-valence effect there is almost no evidence that this causal heterogeneity is ephemeral.

Table 4.

~~Summarized~~ Multilevel Model Output at Joint Analysis of Time 1 and Time 2 Trait-Valence Effects.

Effect	T1 P~~opulation~~arameter Estimates			T1 ~~Population~~ 95% Heterogeneity Interval	
	Mean	SD		2.5%	97.5%
Intercept (or Avg. Level) ~~$(\hat{\mu}_J)$~~ CI$_{95}$	7.05 [7.0, 7.1]	0.19 [0.14, 0.23]		6.67	7.44
Slope (or Causal Effect) CI$_{95}$	-0.14 [-0.13, -0.01]	0.19 [0.08, 012]		-0.50	0.23

Effect	T2 Parameter ~~Population~~ Estimates			T2 95% Heterogeneity Interval~~Population Heterogeneity~~	
	Mean	SD		2.5%	97.5%
Intercept (or Avg. Level) $\hat{\mu}_J$ ~~$(\hat{\mu}_J)$~~ CI$_{95}$	7.00 [6.9, 7.1]	0.22 [0.20, —]		6.58	7.43
Slope (or Causal Effect) CI$_{95}$	-0.19 [-0.17, -0.02]	0.27 [0.13, 0.14]		-0.72	0.34



Formatted: Right, Space After: 0 pt, Line spacing: single

Field Code Changed

Formatted: Right, Space After: 0 pt, Line spacing: single

Formatted: Right, Space After: 0 pt, Line spacing: single

Figure 11. Scatterplot of the predictions of the valence random effects for each participant in the sample at Time 1 and Time 2, with 95% population ellipse.

Given that this level of temporal stability is an estimate from a particular study with a small sample, this result needs to be examined in further studies. It is also important to note that temporal stability may not hold for other heterogeneous experimental effects. The stability may also decline appreciably as the time delay between experimental sessions increases. Nevertheless, this novel approach is useful for assessing the degree to which heterogeneity is attributable to relatively stable tendencies of subjects rather than their temporary psychological and physical states.

Summary

We have stressed the importance of working from the metatheoretical position that experimental effects in psychology are heterogeneous. In Study 1, we showed striking causal effect heterogeneity, which would be completely invisible using standard repeated-measures ANOVA. Not all phenomena show marked heterogeneity, however, and Studies 2 & 3 were intended to distinguish cases where heterogeneity was noteworthy from where it was not. We introduced three criteria, namely the heterogeneity's uncertainty interval, its contribution to model fit, and its size relative to the average causal effect. Further, we demonstrated how the observed heterogeneity in our first example was attributable in part to relatively stable motivational orientations. In Study 4 dataset, we discovered that causal effect heterogeneity showed remarkable stability across two sessions one week apart. Thus, in this example at least, heterogeneity was not just a fleeting effect of subjects' state at the time of the experiment (e.g., fatigue), nor was it an unintended idiosyncrasy of the experimental session. Rather, it reflected something more enduring about how they reacted to the experimental manipulation.

Discussion

We have promoted a way of thinking about experimental effects that is largely absent from experimental psychology, but one that holds much promise: causal effects can vary across individuals in a population (Aim 1). Further, we have shown how using mixed models and graphical displays offer a novel method for experimenters to discover hitherto-unknown heterogeneity in their ~~experimental~~ effects (Aim 2). We have also shown how a concern for causal effect heterogeneity leads to better research practices (Aim 3). When present, causal effect heterogeneity presents opportunities for theory, methods, and research practices in experimental psychology, as we will discuss below.

Opportunities for Theory

Modeling heterogeneity presents an important opportunity for theory development. This need can be especially pertinent if the heterogeneity is sufficiently strong that null effects or reversals are observed. If one assumes that these are not due to failures of experimental control or fleeting states of participants, perhaps the theory needs to accommodate subpopulations that differ in the causal process.

For example, in the face-orientation experiment, it is not clear what factors can explain why some participants respond much faster to upright faces versus upside-down faces, whereas others respond equally fast to each. But knowing that the face orientation effect *is* substantially heterogeneous invites further experiments that manipulate or hold constant explanatory variables such as visual acuity, racial similarity to that of the displayed faces, motivation for the task, etc. For the trait valence experiment, by contrast, theory suggested that the motivational orientation of promotion focus explained heterogeneity in the causal effect, and in fact, our mixed-model results estimated that it accounted for 23% of the heterogeneity.

Sometimes, though, there may be no available explanation for the heterogeneity, and an adequate explanation will require theoretical or methodological breakthroughs that are years or decades away. In this sense, observed heterogeneity can act as a placeholder for future theories and explanatory variables and provide an important qualifier of the generalizability of average causal effects.

Moreover, although we focused on causal effect heterogeneity, the same metatheoretical stance can be applied to other types of relationships between variables to enrich theory. For example, individuals may differ not only in the extent to which they show an experimental effect, but also in the extent to which they vary in mediating processes (Vuorre & Bolger, 2017).

Opportunities for Methods

Although in many cases heterogeneity may reflect meaningful difference between individuals, one spurious source of effect heterogeneity can be uncontrolled variation in experimental procedure across subjects. Some subjects might be run in sessions conducted in summer heat whereas others may not. When multiple experimenters are used in a single experiment, some may put subjects at ease whereas others may not. These are sources that good experimental procedures are meant to minimize. Thus, when experimenters observe effect heterogeneity, it may not be due to true causal differences but rather can be diagnostic of insufficient experimental control. If so, it can lead to salutary revisions in experimental procedures.

Even if procedures are tightly controlled and the tasks and stimuli are valid, experimentalists may view the presence of causal effect heterogeneity as a sign that they should alter their approach. That is, they may change their manipulations or stimulus sets such that the causal effects they produce are homogeneous. Tasks, for example, that evoke different cognitive operations in different subjects, may be replaced with tasks that evoke more homogeneous responses. Such a change might call for alterations in the theory underlying the choice of experimental stimuli. In these cases, the theoretical validity of homogeneity-inducing manipulations or stimuli would need to be demonstrated.

Finally, causal effect heterogeneity can be used to create more efficient experimental designs. If one can understand sources of causal effect heterogeneity (e.g., motivational orientations, as shown earlier) then one could preselect participants for whom an experimental effect is known to be large, thereby allowing one’s sample sizes to be smaller and one’s studies more cost effective (Shrout & Rodgers, 2018). This approach, however, can be criticized for reducing the diversity of samples and limiting generalizability (Tackett et al., 2017).

Implications for Best Research Practices

We view mixed models as an essential tool for analyzing repeated-measures experimental data. Moreover, we believe that repeated-measures ANOVA has outlived its usefulness. We are far from the first to make this point. In 2005, statistician Charles McCulloch wrote an article entitled ‘Repeated Measures ANOVA: RIP?’ urging researchers to switch to the mixed-modeling software that was becoming widespread at the time (McCulloch, 2005). Yet even a cursory look through current journals in experimental psychology will show that repeated-measures ANOVA still predominates in analyses of repeated-measures experiments (as noted in the Introduction). When there are no missing repeated measurements, repeated-measures ANOVA produces correct tests of average causal effects (Maxwell et al., 2018), but we submit that it is a theoretically impoverished account of the data. Even if experimenters wish to focus solely on average causal effects, this approach should ideally be justified by a mixed-model analysis showing that causal heterogeneity is minor and ignorable.

Replication failures, a topic of great current concern (see Shrout & Rodgers, 2018), can be due to failures to take causal effect heterogeneity into account. Replication studies from more heterogeneous populations will be less likely to detect true effects, even if the true average effect size is identical in each population (Bolger & Laurenceau, 2013; Maxwell et al., 2018; Snijders & Bosker, 2011). An important practical implication of greater heterogeneity is that larger sample sizes are needed to maintain adequate power. Because they estimate the size and range of heterogeneity, mixed-model analyses can identify replication failures due to differences in heterogeneity. Power calculations for mixed-model analyses (see e.g., Bolger & Laurenceau, 2013) will allow experimentalists to more effectively plan their future studies. In short, in today's research climate experimentalists can no longer afford to be vague or agnostic about the presence and size of causal effect heterogeneity.

We suspect that causal heterogeneity is present to some degree in all experimental effects, whether these are demonstrated in between- or within-subjects designs. In between-subjects designs, of course, there is no way to assess this heterogeneity without having a manipulation or measured variable that reveals it. But if, as has been argued by Rubin and others (see Imbens & Rubin, 2015; Rubin, 1974; Morgan & Winship, 2014) a *single* causal effect in a between-subjects experiment is equal to an *average* causal effect in a within-subjects experiment, then experimentalists should consider this in interpretations of between-subjects results. Consider the difference between interpreting a causal effect of 0.5 units as uniform across a population versus as an average of hetergeneous causal effects across that population. Thus, even in between-subjects designs, working from the assumption of heterogeneity alters the inferences drawn about the process being studied.

Limitations and Future Directions

We have limited ourselves to models that treat causal effect heterogeneity as a continuous random variable with a parametric distribution, specifically a Gaussian. Generalizations to other continuous distributions are well known and can be implemented in popular software (Gelman & Hill, 2007; Rabe-Hesketh & Skrondal, 2012; Vonesh, 2012). There are reasons to suspect, however, that some forms of heterogeneity are best modeled as categories or classes. An important paper by Lee and Webb

(2005) on cognitive processes treated heterogeneity as involving discrete classes where everyone within a class showed the same causal effect. Models of this sort can be further expanded to include continuous heterogeneity with classes, an approach often called mixture modeling (see, e.g., Bartlema, Lee, Wetsels, & Vanpanel, 2014). Using Bayesian modeling, Haaf and colleagues have proposed a flexible combination of discrete classes with and without further continuous between-subject variation (Haaf & Rouder, 2017, 2018; Thiele, Haaf and Rouder, 2018).

We have also limited ourselves to examining subject-level random effects only. It is well known that mixed models for repeated-measures data should also allow for stimulus-level random effects, so that inferences can be made to a population of stimuli rather than the exact stimuli used in a particular experiment (Clark, 1973). Suitable mixed-models analyses for doing so have been advocated for experimentalists (e.g., Baayen, Davidson & Bates, 2008; Judd, Westfall & Kenny, 2012; Rouder & Lu, 2005). In the Supplemental Material, we present an example of a mixed model with both forms of random effects. None of the results reported in this paper change appreciably when random effects of stimuli are modeled. There are undoubtedly, however, situations in which modeling variability due to stimuli may change causal effect estimates.

Though it is not frequently the case with experimental data, heterogeneous (random) effects in mixed models can sometimes be difficult to estimate using the Frequentist methods used in this paper. Models with Maximum Likelihood estimation of random effects can fail to converge in cases where the effects are not substantial, are poorly estimated, or involve complex models with multiple correlated random effects (see a discussion in Hox, 2012). In these cases, Bayesian estimation will often succeed in producing valid estimates and tests (see Gelman, 2005), although more work is needed to compare random effect estimates obtained using Bayesian vs. Maximum Likelihood methods. For syntax and output for Bayesian versions of our mixed-model analyses, see the Supplemental Material. None of the results presented in this paper were substantially different when Bayesian methods were used.

Also, as noted earlier, examples of sophisticated Bayesian analyses of effect heterogeneity are worth considering. For the interested reader, we recommend a classic paper by Rouder and Lu (2005) and recent work by Haaf and Rouder (Haaf & Rouder, 2017a, 2017b; Rouder & Haaf, 2018). B~~For b~~roader guidance on Bayesian mixed models can be found in ~~, see~~ Gelman & Hill, 2007; Gelman et al., 2013; Kruschke, 2014; Kruschke & Liddell, 2017; Lee & Wagenmakers, 2014; and McElreath, 2016~~)~~. For additional examples of how Bayesian approaches can be used to allow for and investigate heterogeneity in both experimental and non-experimental studies, see papers by Vuorre & Bolger (2017) and Doré & Bolger (2017), respectively.

Perhaps the most sophisticated--and radical--approach to heterogeneity can be found in the work of Molenaar and colleagues (2004; 2009). They question the *a priori* assumption that there are any commonalities in causal processes across subjects. They argue that biological and social units fail to show the thermodynamic property of ergodicity, the property that differences between units at a point in time mirror changes in any unit over time. Non-ergodic processes, they claim, must be examined unit by unit before any inference about commonalities or differences can be made. Thus, their empirical approach is to initially treat each experimental subject as unique and determine with the help of within-subject variation the extent to which subjects can be compared and on what dimensions to do so (see Molenaar, 2004~~2~~; Molenaar & Campbell, 2009).

Finally, we caution that ~~A final point: T~~there are many areas of experimental psychology (beyond the exceptions discussed earlier) where causal effect heterogeneity has simply not been explored. This can be viewed as a ~~caution~~limitation, but it can also be viewed as an opportunity. Consider the vast numbers of existing repeated-measures datasets where heterogeneity has not been

Formatted: Font: Italic

modeled. Without investigators having to collect any additional data, exciting new findings in diverse areas of experimental psychology may be waiting to be discovered.

Conclusions

In order to develop adequate theories of psychological processes, we believe it is advisable to work through all stages of the research process from the assumption that experimental effects are heterogeneous. When planning an experiment, expected causal effect heterogeneity should be taken into account when determining sample size (of subjects and of trials per subject), and when incorporating explanatory variables as additional manipulations or as measured variables.

When analyzing repeated-measures data, mixed models are uniquely able to distinguish true causal effect heterogeneity from spurious sources operating at the subject level such as sampling-error or measurement error. When interpreting and communicating results, the presence or absence of heterogeneity should be featured in causal statements. If heterogeneity is absent, then claims can refer to a universal causal process across the population studied (e.g., the experimental effect had a Cohen's *d* of 0.3). If present, then claims will need to take into account the range of causal effects across a population (e.g., the average person had a Cohen's *d* of 0.3, but some people showed no effect and others showed an effect twice as strong)~~.~~. In either case, these interpretations will be a crucial guide to next steps taken by experimenters in their theory development and in their research plans.

Societies across the globe are becoming more diverse than ever before. Greater diversity will likely lead to greater heterogeneity of experimental effects and require greater richness and realism in our theoretical explanations. Theories and models of experimental data that accommodate heterogeneity are therefore more necessary than ever. Related fields from political science to systems biology to precision medicine have already embraced the notion of causal heterogeneity. We believe it is time for experimental psychology to follow suit.

Formatted: Font: Italic

Formatted: Font: Italic

Context of the Research

This paper grew originally from two ~~a~~ workshops by the first author on Multilevel Models ~~presented~~ at the University of Kent, UK (in 2009 and 2012)- ~~by the first author on Multilevel Models~~ where he showed how these models could be applied to experimental data. ~~for Experimental Data given at the University of Kent, UK (July, 2012)~~ An initial version of the paper was presented at a symposium in honor of philosopher and mathematical psychologist Patrick Suppes at Columbia University (May 2013). Subsequent versions were presented at the Society for Experimental Social Psychology meetings (October, 2015); at the Department of Psychology at Stanford University (February, 2014); at the Department of Psychology at UC Berkeley (April, 2014); at the Department of Psychology at Columbia University in 2015, 2016, and 2017; and at a conference on *Evidence: An Interdisciplinary Conversation* at Columbia University (April, 2017). The authors are indebted to the many comments and suggestions made by participants at these events, and to comments on an earlier draft of the paper by Patrick Shout of New York University and Megan Goldring of Columbia University. We also thank Asael Sklar and colleagues for contributing the datasets ~~for~~ examined in Studies 2 and 3, and Abigail Scholer and colleagues for contributing the dataset examined in Study 4.

Some of the ideas in the paper draw on earlier work by the first author on personality-based causal heterogeneity in stress and coping processes (Bolger, 1990; ~~reactivity by~~ Bolger and Schilling ~~(~~1990; Bolger & Zuckerman, 1995; Bolger & Romero-Canyas, 2007)~~)~~; from work on how to incorporate causal heterogeneity in analyses of intensive longitudinal data (~~a book by~~ Bolger, Davis & Rafaeli, 2002; Bolger and Laurenceau, ~~(~~2013) ~~highlighting causal heterogeneity in models of intensive longitudinal data~~; and from a ~~broader~~ program of research on social support ~~adjustment~~ processes in ~~close relationships~~ in experimental and naturalistic settings (Bolger, Zuckerman & Kessler, 2000; Bolger & Amarel, 2007) . ~~A current direction in this line of work examines how heterogeneity in close relationship processes changes across the lifespan.~~

References

Akdoğan, B., & Balcı, F. (2017). Are you early or late?: Temporal error monitoring. *Journal of Experimental Psychology: General*, *146*(3), 347.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The Economic Journal, 114*, C52-C83.

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.

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology, 59*, 132-150. doi: https://doi.org/10.1016/j.jmp.2013.12.002

Bolger, N., & Laurenceau, J. P. (2013). Intensive longitudinal methods. *New York, NY: Guilford*.

Brand, J. E., & Thomas, J. S. (2013). Causal Effect Heterogeneity. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 189-213). Dordrecht: Springer Netherlands.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Thousand Oaks, CA: Sage.

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

Darwin, C. (1859). *On the origin of species by means of natural selection*. London: John Murray.

Doré, B., & Bolger, N. (2017). Population- and Individual-Level Changes in Life Satisfaction Surrounding Major Life Stressors. *Social Psychological and Personality Science*. doi: 10.1177/1948550617727589

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134-140.

Estes, W. K., & Todd Maddox, W. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review, 12(*3), 403–408.

~~Fisher, C. I., Hahn, A. C., DeBruine, L. M., & Jones, B. C. (2015). Women's Preference for Attractive Makeup Tracks Changes in Their Salivary Testosterone. *Psychological Science, 26*, 1958-1964. doi: 10.1177/0956797615609900~~

Formatted: Indent: Left: 0", First line: 0"

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935). *The Design of Experiments.* Oxford: Oliver & Boyd.

Fisher, C. I., Hahn, A. C., DeBruine, L. M., & Jones, B. C. (2015). Women's Preference for Attractive Makeup Tracks Changes in Their Salivary Testosterone. *Psychological Science, 26*, 1958-1964. doi: 10.1177/0956797615609900

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Science, 115, E6106-E6115.* doi: 10.1073/pnas.1711978115

Fitzmaurice, G. M., & Molenberghs, G. (2009). Advances in longitudinal data analysis: An historical perspective. In G. M. Fitzmaurice, M. Davidian, G. Verbeke & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 3-27). Boca Raton: CRC Press.

~~Förster, J., Higgins, E. T., & Bianco, A. T. (2003). Speed/accuracy decisions in task performance: Built-in trade-off or separate strategic concerns?. *Organizational Behavior and Human Decision Processes, 90*(1), 148-164.~~

Gelman, A. (2005). Analysis of variance: Why it is more important than ever. *The Annals of Statistics, 33*, 1-31.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, Fla.: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Haaf, J. M., & Rouder (2017). *Developing constraint in bayesian mixed models*. Manuscript under review.

~~Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39*, 101-117.~~

~~Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods, 39*, 723-730. doi: 10.3758/bf03192962~~

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods, 22*, 779-798. doi: 10.1037/met0000156

Haaf, J. M., & Rouder, J. N. (2017). Some do and some don't? Accounting for variability of individual difference structures. doi: https://doi.org/10.17605/OSF.IO/ZWJTP

Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930*. New York: Wiley.

~~Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science, 26, 10-15.* doi: 10.1177/0963721416666518~~

Higgins, E. T. (1998). Promotion and prevention: Regulatory focus as a motivational principle. *Advances in Experimental Social Psychology*, *30*, 1-46.

Higgins, E. T., Friedman, R. S., Harlow, R. E., Idson, L. C., Ayduk, O. N., & Taylor, A. (2001). Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology*, *31*(1), 3-23.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39*, 101-117.

~~Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods, 39*, 723-730. doi: 10.3758/bf03192962~~

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics, 7*(1), 443-470.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. New York: Cambridge University Press.

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)*, 54-69.

Kenny, D. A., & Judd, C. M. (In press). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Science*.

Kruglanski, A. W., Thompson, E. P., Higgins, E. T., Atash, M., Pierro, A., Shah, J. Y., & Spiegel, S. (2000). To" do the right thing" or to" just do it": Locomotion and assessment as distinct self-regulatory imperatives. *Journal of Personality and Social Psychology, 79*(5), 793-815.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. San Diego, CA: Academic Press.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. [journal article]. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-016-1221-4

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). Boston: McGraw-Hill/Irwin.

Leary, M. R., & Baumeister, R. F. (2000). The nature and function of self-esteem: Sociometer theory. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology, 32,* 1-62. San Diego, CA: Academic Press.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605-621.

Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods, 39*, 723-730. doi: 10.3758/bf03192962

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Mahwah, NJ: Erlbaum.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models* (2nd ed.). New York: Wiley.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples from R and Stan*. CRC Press.

Miller, J., & Schwarz, W. (2018). Implications of individual differences in on-average null effects. *Journal of Experimental Psychology: General, 147*, 377-397. doi: 10.1037/xge0000367

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2(4),* 201-218.

Molenaar, P. C. M., & Campbell, C. G. (2009). The New Person-Specific Paradigm in Psychology. *Current Directions in Psychological Science, 18*, 112-117.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*, 103-123. doi: 10.3758/s13423-015-0947-8

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). New York: Cambridge University Press.

OECD. (2017, 20170923). Program for International Student Assessment. Retrieved October 23rd, 2017.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8-13.

Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale [On man and the development of his faculties, or Essay on social physics]*. Paris: Bachelier.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rabe-Hesketh, S., & Skrondal, A. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural* equation models. New York: Chapman and Hall/CRC.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata: Categorical responses, counts, and survival* (3rd ed. Vol. 2). College Station, Tex.: Stata Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science, 1*, 19-26. doi: 10.1177/2515245917745058

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic bulletin & review*, *12*(4), 573-604.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*, 356-374. doi: https://doi.org/10.1016/j.jmp.2012.08.001

Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). Acceptance and commitment therapy. *Measures package, 61*, 52.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66(5),* 688-701.

Scholer, A. A., Ozaki, Y., & Higgins, E.T. (2014). Inflating and deflating the self: Sustaining motivational concerns through self-evaluation. *Journal of Experimental Social Psychology, 51,* 60-73.

Shrout, P., Rodgers, 2018 [preselect participants for whom an experimental effect is known to be large, thereby allowing one's sample sizes to be smaller and one's studies more cost effective]

Sklar, A. Y., Goldstein, A. Y., Abir, Y., Dotsch, R., Todorov A., & Hassin, R. R. (2017) *Non-conscious speed: A robust human trait*. Manuscript under review.

Sklar, A. Y., Levy, N., Goldstein, A., Mandel, R., Maril, A., & Hassin, R. R. (2012). Reading and doing arithmetic nonconsciously. *Proceedings of the National Academy of Sciences, 109(48)*, 19614-19619.

Snijders, T., & Bosker, R. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. Boca Raton, FL: Taylor & Francis/CRC Press.

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., ... & Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*(5), 742-756.

Thiele, J. E., Haaf, J. M., & Rouder, J. N. (2017). Is there variation across individuals in processing? Bayesian analysis for systems factorial technology. *Journal of Mathematical Psychology, 81*, 40-54. doi: https://doi.org/10.1016/j.jmp.2017.09.002

Western, B. (1998). Causal heterogeneity in comparative research: A bayesian hierarchical modelling approach. *American Journal of Political Science, 42(4)*, 1233-1259.

Whitsett, D. D., & Shoda, Y. (2014). An approach to test for individual differences in the effects of situations without using moderator variables. *Journal of Experimental Social Psychology, 50,* 94-104.

Vonesh, E. F. (2012). *Generalized linear and nonlinear models for correlated data theory and applications using SAS*. Cary, NC: SAS Institute, Inc.

Vuorre, M., & Bolger, N. (2017). Within-subject mediation analysis for experimental data in cognitive psychology and neuroscience. *Behavior Research Methods*. doi: 10.3758/s13428-017-0980-9

Xie, Y. (2013). Population heterogeneity and causal inference. *Proceedings of the National Academy of Sciences, 110(16)*, 6262-6268.

Yamaguchi, S., Greenwald, A. G., Banaji, M. R., Murakami, F., Chen, D., Shiomura, K., Kobayashi, C., Cai, H., & Krendl, A. (2007). Apparent universality of positive implicit self-esteem. *Psychological Science*, *18*(6), 498-500.

Formatted: Font: 11 pt

Formatted: Line spacing: Double

Formatted: Font: Calibri, 11 pt

Formatted: Font: Calibri, 11 pt

Formatted: Indent: First line: 0.38", Line spacing: Double, Tab stops: 3.55", Left

Formatted: Font: Calibri, 11 pt

[1] When there are no missing repeated measurements, and there are mutiple trials for each cell of the within-subjects design, results of repeated-measures ANOVA can be used to estimate effect heterogeneity (see Keppel & Wickens, 2004; Maxwell & Delaney, 2017). When faced with missing trials in a cell, however, researchers often aggregate over trials and use cell-means as input to repeated-measures ANOVA software, thereby making it impossible to assess effect heterogeneity.

[2] The paper by Akdogan & Balci (2017) did compute individual-specific effects and report the range and standard deviation of these effects. However, these effects do not appear to have been derived from mixed models, but rather from models run on data from each individual.

[3] Our primary construct of interest was regulatory focus, given prior work showing the implications of promotion focus and endorsement of positive words using this paradigm (Scholer et al., 2014). We also included measures of two other constructs that could explain some of the variability in reaction times to endorse positive and negative traits: regulatory mode orientations (locomotion and assessment) and self-esteem. Regulatory mode orientations were measured using the Regulatory Mode Questionnaire (Kruglanski et al., 2000), which consists of 12 items measuring locomotion and 12 items measuring assessment. Self-esteem was measured using Rosenberg's (1965) 10-item measure. However, as determined *a priori* based on earlier work, we were primarily interested in promotion focus. Thus, measures of regulatory mode and self-esteem will not be discussed further. The promotion focus subscale of the Regulatory Focus Questionnaire consists of six items rated along a scale ranging from 1 (*never or seldom*) to 5 (*very often*).

Formatted: Font: 10 pt

[4] For more information about treating time or temporal ordering of stimuli as random effects, see Chapter 4 of Bolger & Laurenceau (2013). For more on treating stimuli as random effects, see Judd, Westfall, & Kenny (2012).

[5] Because our dataset is unbalanced (only self-relevant traits were analyzed), a repeated-measures ANOVA would require that the RTs for trials (stimuli) within valence condition be aggregated, leading to just two observations per subject. Such a data set would obscure any causal effect heterogeneity.

Formatted: Font: 10 pt

Formatted: Normal

[6] Values in all tables draw on results from models run in R. SPSS results may vary slightly due to rounding. We ~~Due to our focus on causal effect heterogeneity, we~~ focus mostly on parameter estimates and confidence intervals in this paper;~~.~~ a~~A~~dditional summaries such as ~~*statistics* (~~*t*-tests and *p*-values~~s)~~ can be found in the Supplemental Material.

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: 10 pt

Formatted: Font: Italic

[7] Note that the 95% confidence refers not to this specific interval but to the long-run performance of the *procedure* of creating confdence intervals in hypothetical replications of the study (see Morey et al., 2016). Nonetheless, this specific interval is evidence as to the location of the population effect, even if it does not have a probability interpretation (Mayo, 2018). Readers wishing to have a probability interpretation of parameter intervals, should examine the Bayesian versions of all analyses in the Supplementary Materials. With noninformative priors on all model parameters (see Gelman et al., 2013), the 95% posterior credibility interval for the equivalent effect ranges from -0.21 to -0.12 logRT units. In general, we find the Frequentist and Bayesian estimates and intervals to be very similar numerically, and where they diverge substantially we note this in the body of the paper.

[8] The right-most panel in Figure 2 would suggest that the subject who showed the weakest valence effect also endorsed relatively few positive words. We ran additional versions of this analysis to rule out the possibility that number of words endorsed or asymmetry in endorsement played a role in our results. See Supplement 5.

Formatted: Font: 10 pt

~~[9] Note that one additional participant could not be included in this visualization because this person's observed difference could not be computed as only one negative trait was endorsed.~~

[10] Note that one additional participant could not be included in this visualization. This is because ~~because~~ this person's observed difference could not be computed as only one negative trait was endorsed.

[11] More details about the study methods and data processsing (e.g., exclusion criteria) can be found on pages 119614 and 9617-8 of Sklar et al. (2012). The experiment also involved a between-subjects manipulation of presentation time (1700 ms vs. 2000 ms). For simplicity ~~Because of our focus on causal effect heterogeneity,~~ we do not include presentation time in our analyses. In other words, the analysis presented here examines the congruence effect across both presentation time conditions. Including presentation time in the model had a negligible effect on the heterogeneity results~~The conclusions regarding heterogeneity, or lack thereof, in the math priming effect remain unchanged when accounting for presentation time in the model~~. Also note that the original Sklar et al. (2012) paper analyzed data in milliseconds, but the analysis presented here is in log milliseconds.

[12] Some work (see papers by Haaf, Rouder, and colleagues) suggests a potential relationship between effect size and the amount of variation in people's responses to an experimental manipulation; they point to the case where effect reversals are not justified by theory or logic. In such cases, one would expect a floor effect on the distribution of effects, which would imply that smaller average effects would be accompanied by smaller heterogeneity. If effect reversals are to be expected for some proportion of the population, then one would not expect effect size and heterogeneity to be proportional (see Miller & Schwarz, 2018, for a relevant discussion).

[13] ~~We acknowledge that~~ P~~p~~revention focus is also an important motivational orientation, and given that it was measured in the study~~.~~ ~~, and that "best practices" in motivation research involve including both promotion and prevention as simultaneous predictors in all analyses. As such, w~~wWe also performed a analysis ~~version of this analysis~~ that included valence, promotion focus, prevention focus, and all possible interaction terms. There was an interaction of valence and prevention focus, but it was only marginally signficant. Moreover, ~~However,~~ given that the main effect of valence and the promotion by valence interaction were ~~described in this section were~~ essentially unchanged~~the same as the results presented~~~~,~~, for brevity we ~~describe~~ presented the simplified model only~~without prevention focus in the main text~~. ~~Results for the more complex analysis are available in the Supplemental Materials.~~ ~~for ease of presentation.~~

[14] In the Scholer et al. (2014) paper, there was a between-subjects experimental induction of Regulatory Focus (promotion or prevention) prior to the Time 2 trait valence task. A version of the analysis with this between-subjects manipulation included resulted in minimal changes to the results. This makes sense when one considers that due to random assignment each participant had an equal chance of being assigned to the promotion or prevention induction, regardless of the size of their Time 1 trait valence effect.~~Although we did not control for induction in the temporal stability analysis presented here, doing so would not change our inferences regarding temporal stability. Due to random assignment, each participant had an equal chance of being assigned to the promotion induction and prevention induction, regardless of the size of their trait valence effect.~~
~~Indeed, running a version of the analysis with this between-subjects manipulation included did not change the results appreciably, nor did it alter our conclusions regarding the stability of between-subject heterogeneity across timepoints.~~

[15] Due to the additional complexity involved in this analysis, statistical code and output are provided in R only.

[16] We also performed ~~an~~ additional analyses to better understand temporal stability in the trait valence effect. In one ~~version of this~~ analysis, ~~in which~~ we ran ~~a~~ t-tests treating each participant as their own sample~~, just like the approach used in Study 1~~. ~~We then computed the correlation coefficient for the observed differences in reaction time as a function of valence at Time 1 and Time 2. We found that the observed differences were correlated at the two time points but that this correlation was weaker, *r* = .80.~~In another analysis, we examined temporal stability using Bayesian estimation. In both cases, the correlations between Time 1 and Time 2 subject-specific effects was noticeably lower, in the .7-.8 range~~were not as high~~. See Supplement 9. ~~We also performed an additional version of this analysis in which we ran a t-test treating each participant as their own sample, just like the approach used in Study 1. We then computed the correlation coefficient for the observed differences in reaction time as a function of valence at Time 1 and Time 2. We found that the observed differences were correlated at the two time points but that this correlation was weaker, *r* = .80.~~

Formatted: Font: 10 pt

Formatted: Font: 10 pt

Formatted: Font: 10 pt

Formatted: Font: 10 pt

Formatted: Font: 10 pt

Formatted: Font: 10 pt

Formatted: Font: 10 pt

Formatted: Font: 9 pt