# Causal Heterogeneity: Opportunities and Pitfalls in Experimental Data

# Abstract

All experimenters know that human and animal subjects show natural variability and do not respond to experimental treatments in a uniform way. It is common to view this causal heterogeneity as a source of error, and in traditional between-subjects designs it is often modeled in this way. In within-subjects designs, however, where causal inference involves comparing subjects to themselves, such heterogeneity can be examined directly and can yield important theoretical and methodological insights. We argue that many existing datasets contain such untapped opportunities, and using datasets from cognitive and social psychology, we show how to avail of them. This guidance also applies to those planning to conduct such studies. We also show the potentially grave consequences of failing to take causal heterogeneity into account and the tremendous value that can be gained from exploring it.

# Introduction

All organisms show intrinsic heterogeneity, that is, variation in their structure and function that cannot be simply attributed to measurement error. In psychology it is common to conceptually consider such variability, for example, in terms of individual differences. Within species in nature, there is heterogeneity in basic phenotypic features such as size, shape, color, symmety. This is true for microorganisms, plants, animals. Humans are no exception. It can be argued that this fact lies at the foundation of modern statistics. At least since the time of Galton, intra-specific variability has been recognized. In Fisher’s landmark book, *Statisical Methods for Research Workers*, heterogeneity is the background used for determining whether an experimental effect is likely to be real or not. His ANOVA model compares mean differences between treatments and control to what would be expected if that variation had occured from background heterogeneity in experimental units.

It is typical practice in experimental psychology to view heterogeneity as unwanted noise in statistical models for experimental data. The focus is on the average not the variability. If that variability can be attributed to measurement error, irregularities in experimental procedure, or temporary states of participants (e.g., being hungry or tired), this approach makes sense. If, however, it captures more eduring differences between subjects in their responses to the experimental stimuli, then this approach leads to lost opportunities for understanding the phenomemon being investigated. For example, finding heterogeneity in how individuals respond to a particular stimulus can help elucidate the mechanisms behind the experimental effect. In addition, ignoring heterogeneity can lead to spurious findings, which in turn lead to distorted literatures and failures to replicate.

In typical between-subjects experimental designs, where one observation is obtained on each subject, causal heterogeneity cannot be distinguished from measurement error and other process-irrelevant variability in responses, even though it may be present. In within-subjects designs, however, in which causal inference involves comparing subjects to themselves in other experimental conditions, the existence and size of heterogeneity can distinguished from other sources of variance. Our goal in writing this paper is to show how examiniation of heterogeneity can be accomplished. The basic statistical model we will use is in essence repeated-measures ANOVA, but implemented as a mixed model (also known as a multilevel or mixed-effects model). The value of a mixed model analysis for understanding causal heterogeneity will become clear in following section.

I think an overview is needed here. What experiments are we going to talk about and why, what will we discuss when, etc. and then maybe a discussion of the overall advantages that await those who read you and implement you. these are things that are clear to you (and now to me too), but are less clear for simple-minded experimentalists.

# Uncovering Causal Heterogeneity: The Example of Stimulus Valence Effects on RT

Our first example dataset comes from a 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 described them as reaction time was measured. The research question we are interesed in is whether there are reaction time differences in endorsed self-descriptions according to whether their valence is positive or negative. A straightforward prediction is that participants will be faster to endorse positive as opposed to negative self-descriptions. We chose this straightforward hypothesis not because of its scientific interest but because we wanted to examine heterogeneity in a robust effect.

## Methods

Participants

Seventy-five students from Columbia University  completed the study for either 1 course credit or $5.  Thirteen were excluded for failing an attention check. The final N was 62.

Procedure

After giving their consent, participants were led into individual cubicles to begin the experimental task, administered on a computer with PsychoPy software (cite). Participants completed the Regulatory Focus Questionnaire (Higgins et al., 2001), additional individual difference measures that we will not be discussing further, and general demographics.  After participants answered these questionnaires, they completed a computerized task measuring the trait valence effect. Finally, they were debriefed and thanked.

Measures

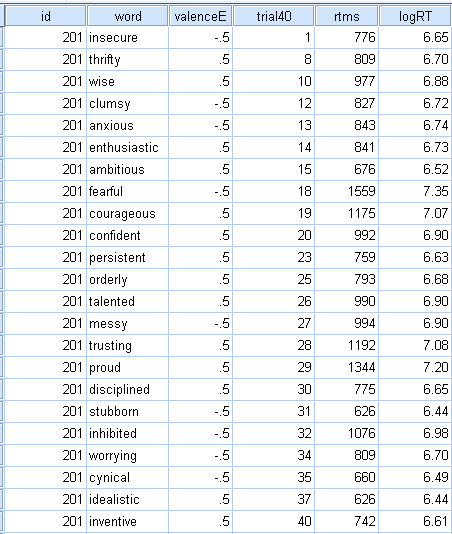
*Measure of trait valence effect*.   Each trial began with a fixation point that appeared for 1 second, followed by a personality trait word.  The participants’ task was to judge whether they possessed the trait or not, as quickly as possible, by pressing a designated key on the keyboard.  If the answer was yes, they pressed the “P” key with the right index finger.  If the answer was no, they pressed the “Q” key with the left index finger. When the response was made, the trait word disappeared, and 2 seconds later the next trial began.  The first 6 trials served as the practice phase, followed by 34 experimental trials.

Each personality trait appeared only once, in a distinct random order for each participant. The computer recorded the response latency (i.e., the time elapsed between the appearance of trait word and the key-press by participants) as well as the yes/no response.

## Mixed model analysis

As is common with reaction time data, we use the natural log transformation to remove skewness. Only trials containing words endorsed as self-relevant were included in the analyses. To examine our hypothesis of stimulus valence effects on logRT, we will primarily work with a statistical model where stimulus valence is the single predictor and reaction time in log-milliseconds is the outcome. This basic model can be used to develop the main points regarding heterogeneity of causal effects. Specifically, it will allow us to examine whether people do, on average, respond faster to positive than negative self-relevant traits, while also allowing us to examine the variability in this effect. Later we will consider the need to (a) add additional information about the temporal ordering of the stimuli and (b) allow for stimuli and temporal order to be treated as random effects. All datasets, together with syntax and code, are available at the following URL. We begin these analyses in SPSS and R.

If we examine in Table 1 a portion of the dataset we can see that each RT observation for each subject is given a separate data line. Each line contains columns for the subject’s unique indentifier (id), the exact stimulus word used (such as “confident” and “anxious”), the valence of each word (valenceE, coded -0.5 if negative and 0.5 if positive), and the temporal order in which each word was presented to this subject (trial40, e.g., 1st, 15th, etc.). Recall that because stimulus ordering was randomized, each subject encountered a different order. Finally there are columns for the reaction time to the stimulus word in milliseconds (rtms) and in log milliseconds (logRT).



Our model specifies that in the population, a person’s reaction time can be different for positively valenced and negatively valenced traits, and it allows this difference to vary across persons. In other words, our model accounts for the fact that some people tend to respond faster to positive than negative traits, while others tend to respond to them at relatively equal speeds or even respond faster to negative traits. This model is parametrically identical to a standard repeated-measures model with a single within-subjects factor with 20 stimulus replications within each of two factor levels (see Kirk, Winer, Maxwell & Delaney for descriptions repeated-measures ANOVA). Here we will use a mixed or or multilevel modeling approach (McCulloch, Searle & Niewhuis, 2012) that unlike repeated-measures ANOVA will reveal the existence and size of causal heterogeneity.

The model is composed of three equations. The first specifies that within-subject variability in trial-level logRT scores is a function of stimulus valence:

 (1)

In Eq. 1, the logRT observed for subject *j* (1, 2, …62) for the stimulus in trial *i* is specified to be normally distributed with a subject-specific mean and a common standard deviation. The mean is a linear function composed of , a grand-mean level effect that varies across subjects j and , a valence effect term where the valence coefficient also varies across subjects j (stimulus valence  is coded -0.5 if the stimulus is negative and 0.5 if the stimulus is positive). The common standard deviation is , which refers to the scatter of residual RT scores around the specific  for each subject. In this example, subjects with larger grand-means  tend to respond to traits more slowly than subjects with smaller grand means. Subjects with negative  values tend to respond relatively faster to positive than negative traits, whereas subjects with postive  values show the opposite tendency.

The second and third specify the distribution of heterogeneity in the within-subjects effects:

 (2)

 (3)

In Eq. 2, the subject-specific grand-mean effects are specified to be normally distributed with a mean  that represents the population average grand-mean of logRT, and a standard deviation  that represents heterogeneity (i.e., between-subject variability) in grand-means. In Eq. 3, the subject-specific valence effects are specified to be normally distributed with a mean  that represents the population average valence effect, and a standard deviation  that represents heterogeneity (i.e., between-subject variability) in valence effects.

The syntax requred to estimate the model in SPSS is:

MIXED logRT WITH valenceE

/FIXED=valenceE

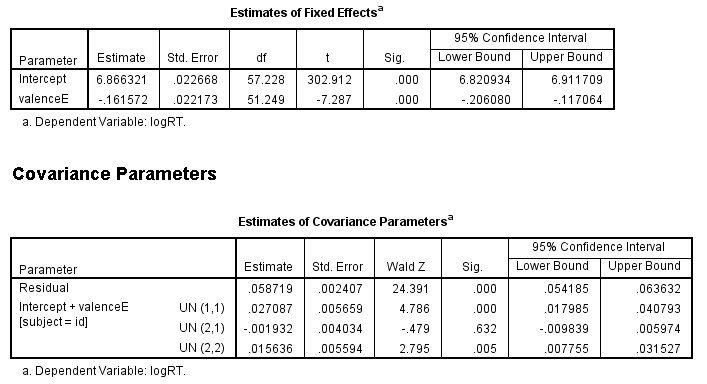
/RANDOM=INTERCEPT valenceE | SUBJECT(id) COVTYPE(UN).

/PRINT=SOLUTION TESTCOV

For R, it is:

Out <- lmer(logRT ~ valenceE + ( 1 + valenceE| id ), data = study1)

Both produce identical results. SPSS:



R:

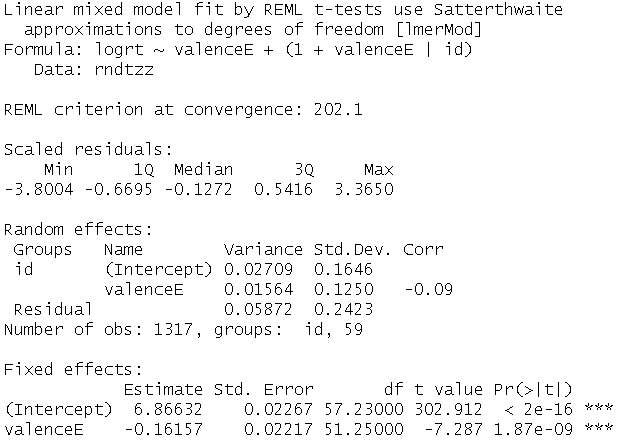


Table 1 summarizes the key estimates of interest. These are the parameters from Equations 2 and 3. The grand-mean logRT () for the typical person is 6.87 units. The heteroegeneity in grand means, as reflected in their SD (), is 0.16 units, which implies that the 95% population range is from 6.6 to 7.2 units. The valence effect for the typical person () is -0.16 logRT units faster (approximately 160 ms) at responding to positively valenced words than to negatively valenced words. The heterogeneity in this effect, as reflected by its SD () is 0.13 units, which implies that the 95% population range is from -0.41 to 0.09. This level of heterogeneity suggests that a one-size fits all view of valence effects that one gets from the fixed effect is quite misleading.

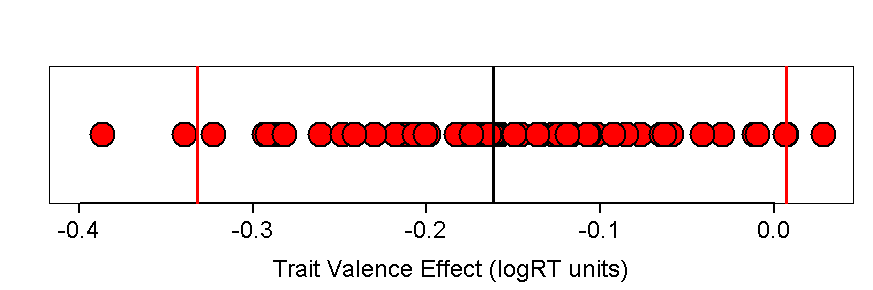
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Effect | Population Estimates | |  | Population Heterogeneity | |
|  | Mean | SD |  | 2.5% | 97.5% |
| Level () | 6.87 | 0.16 |  | 6.6 | 7.2 |
| Valence () | -0.16 | 0.13 |  | -0.41 | 0.09 |
|  |  |  |  |  |  |
| CI95 (Level) | [6.8, 6.9] | [0.13, 0.20] |  |  |  |
| CI95 (Valence) | [-0.21, -0.12] | [0.08, 0.18] |  |  |  |

The model not only provides estimates of population parameters. It can also be used to predict the valence effect for each person in the sample. These are called Empirical-Bayes predictions or Best Linear Unbiased Predictions (BLUPs). We display these in two ways. First we show them as a stripchart in Figure x. The large red dots represent each individual in the sample The black line in the center is the fixed effect and the red lines at the edges lines are the 2.5 and 97.5 percentiles for the sample. This illustrates the substantial causal heterogeneity observed: Although the average participant is -.016 logRT units faster at responding to positive vs. negative traits (A), a participant at the lower bound show an effect twice as large, -0.33 units (B), and a participant at the upper bound show no difference in logRT (C).

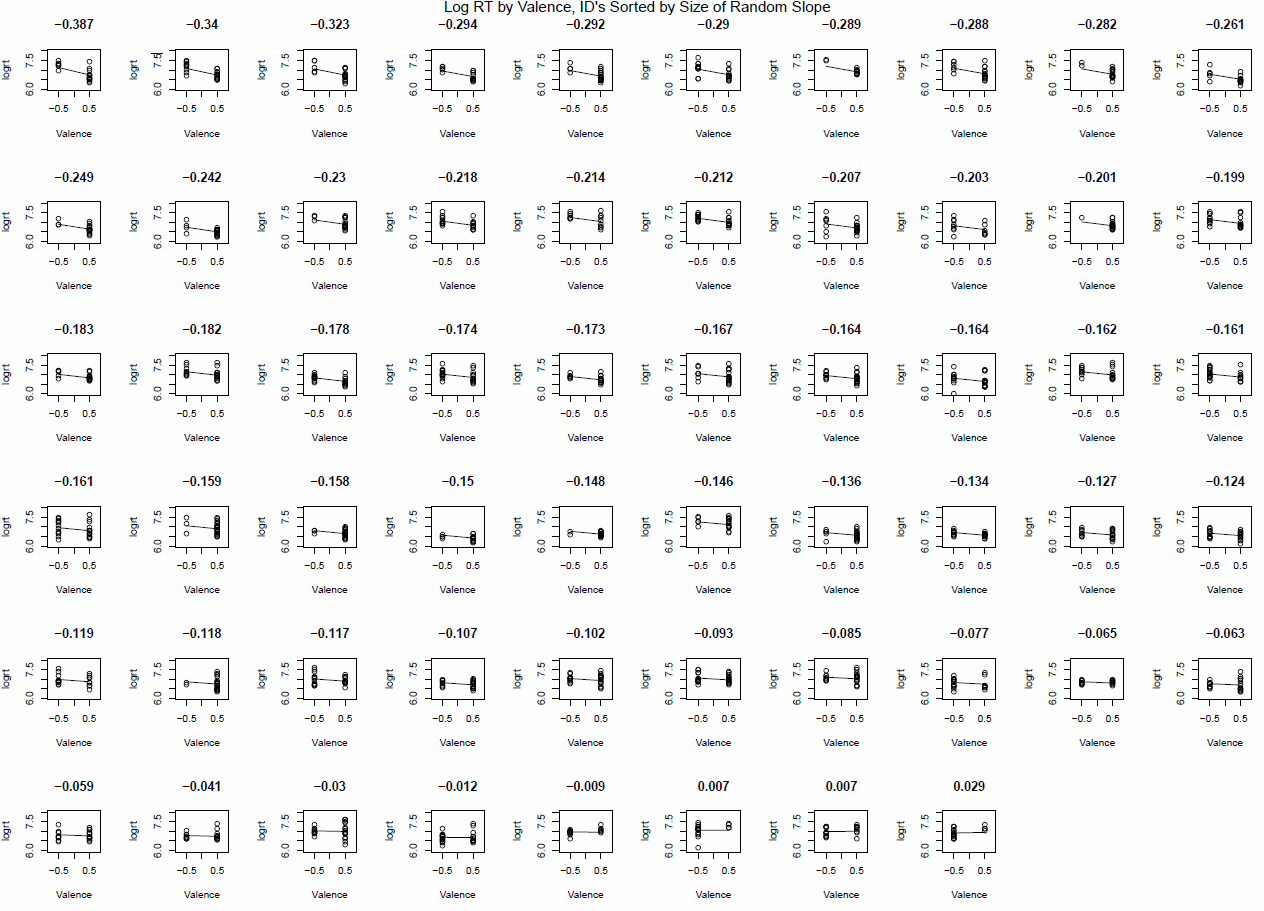
C

B

A



The second way is to overlay the predictions from our model on the observed data. Figure 1 is a panel plot showing each subject’s data for logRT as a function of valence together with the model-predicted values for the subject. Each fitted line correspond to a valence-effect data point in the stripchart above. Like the stripchart the panels are ordered by the size of the model-predicted valence effect. Overall, the reader should see that the predictions capture valence effects that are visible in each participant’s raw data, and that these valence effects show marked heterogeneity. The subject in the upper left-hand corner shows a difference that is approximately 2.4 times larger than the subjects in the middle of the panels ( ≈ -0.39/-0.16). This heterogeneity is evident in the raw data and in the model predictions. To reiterate, although the fixed effect from our model revealed that the typical person was about 0.16 logRT units faster to endorse positively valenced traits as self-relevant compared to negatively valenced traits, it is apparent from this visualization that sample members differ in the magnitude and in some cases even the direction of this effect.



The reader will see the extra information contained in the mixed model results compared to repeated-measures ANOVA. Allowing for heterogeniety allows to to determine a single valence effect is sufficient to characterize the experimental phenomenon, or whether person-specific effects are needed. In the current dataset, they are needed, and the observed heterogeneity calls for theory that allow one to explain their existence and magnitude. We will consider the question of explaining heterogeneity later in the paper. But even in the absence of explanatory variables, the causal heterogeneity is fundamental to any accurate account of the experimental results.

## Why not repeated measures ANOVA?

Since we describe the model above as parameterically identical to a repeated-measures ANOVA, the reader may naturally wonder why we could not use traditional repeated-measures software. On a practical level, one reason is that we do not have a balanced design. We are examining logRT to traits that subjects say are self-descriptive and as can be seen from Figure 1, fewer negative traits than positive traits are chosen by most subjects. Had we not selected on self-descriptive traits, we would still be unable to use repeated-measures ANOVA because 18 of 62 participants had missing repeated measurements. This is not uncommon in designs with many repeated measurements and it is the main reason researchers who use repeated-measures ANOVA tend to aggregate their data before analyzing them. This as we discuss below results in incorrect inferences. Even without aggregation, however, the repeated-measures ANOVA is deficient on conceptual grounds: It does not provide estimates of heterogeneity of effects (i.e., , ).

## The aggregation approach

When datasets are unbalanced--as is the case with Study 1--one common solution is to aggregate the data within each subject so that only cell means on the repeated factor are used. In this way, the data can be analyzed using standard repeated measures ANOVA. For Study 1 this would mean that we can use data from all but the two participants who chose no negatively-valenced traits.

To conduct a repeated-measures ANOVA on aggregated data in SPSS:

GLM logrtneg logrtpos

/WSFACTOR=valence 2

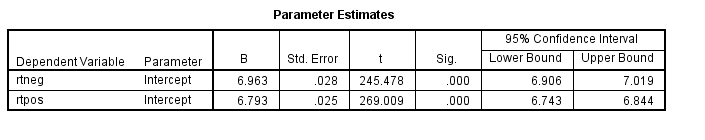
/METHOD=SSTYPE(3)

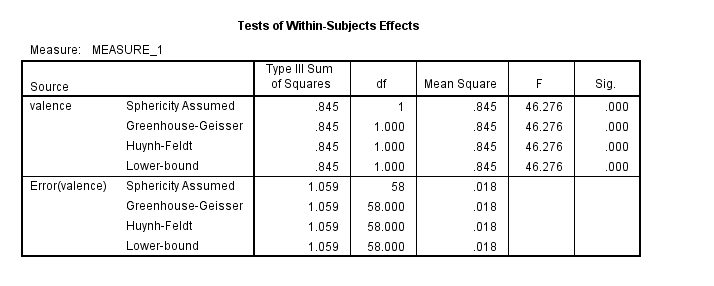
/WSDESIGN=valence.

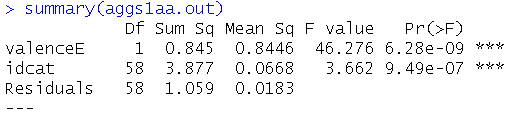
In R:

aggs1aa.out = aov(logrt ~ valenceE + idcat, data=aggs1aa)

The outputs are:







|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Effect | Population Estimates | |  | Population Heterogeneity | |
|  | Mean | SD |  | 2.5% | 97.5% |
| Level () | 6.88 | 0.16 |  |  |  |
| Valence () | -0.17 | -- |  |  |  |
|  |  |  |  |  |  |
| CI95 (Level) | [6.8, 6.9] |  |  |  |  |
| CI95 (Valence) | [-0.22, -0.12] |  |  |  |  |

As we can see, aggregation over the stimuli within the negative and positive conditions has not changed the results appreciably. In fact, had the original data been balanced, the estimates and tests would have been identical. This follows because aggregation affects the signal and the noise in the ANOVA in the same way, leaving the ratio of the two unchanged. What has happened to the heterogeneity in this case? We define heterogeneity as the true differences between subjects in the size of the trait-valence effect. It is still present in the data but, using this analytic approach, it cannot be distinguished from error variance due to the specific stimuli in each valence condition. In other words, we cannot tell whether the variability in responses is due to differing tendencies between participants, or simply due to error. We see this as a major limitation of the aggregation approach as it overlooks important information about the effect.

## Failing to model the heterogeneity

Although using mixed-model software is essential to estimating and displaying heterogeneity in causal effects, it requires care. Some researchers have made the mistake of omitting heterogeneity in their models. The most notable in recent years has been a paper by Fisher et al. 2015 that was retracted from *Psychological Science*. In this case, although the mixed model included the correct within-subjects factors, it did not allow for possibile heterogeneity in these effects. This led to a severe upward bias in test statistics, a bias that, when eliminated, rendered the authors’ findings inconclusive. Let us examine for our example dataset the consequences of omitting heterogeneity effects. In the following analyses, we retain the random intercept term for each person (allowing each person to differ in how quickly they tend to respond to traits overall), but remove the random slope component thus imposing the same effect of valence on each participant.

SPSS:

MIXED logrt WITH valenceE

/FIXED=valenceE | SSTYPE(3)

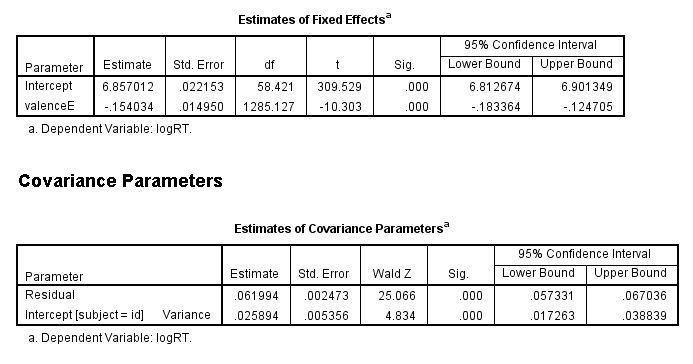
/METHOD=REML

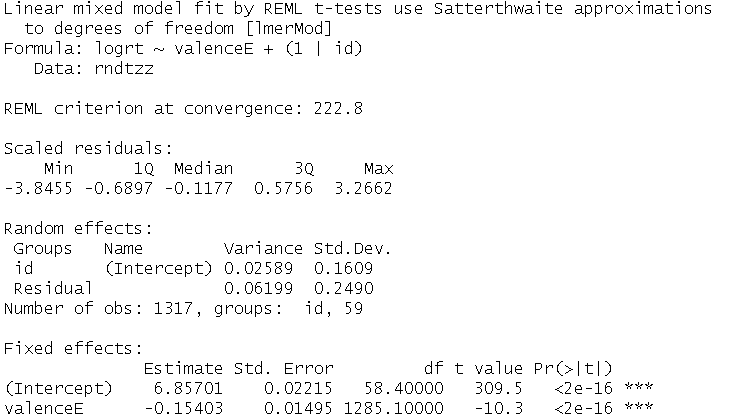
/PRINT=G SOLUTION TESTCOV

/RANDOM=INTERCEPT | SUBJECT(id) COVTYPE(UN).

R:

mle3aaa <- (lmer(logrt ~ valenceE + (1| id), data=rndtzz))





|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Effect | Population Estimates | |  | Population Heterogeneity | |
|  | Mean | SD |  | 2.5% | 97.5% |
| Level () | 6.86 | 0.16 |  | 6.5 | 7.2 |
| Valence () | -0.15 | -- |  | -- | -- |
|  |  |  |  |  |  |
| CI95 (Level) | [6.8, 6.9] | [0.13, 0.20] |  |  |  |
| CI95 (Valence) | [-0.18, -0.12] | -- |  |  |  |

Again the results for SPSS and R are identical. What has changed in these results is the standard error and t value for the valence effect. In the random intercepts and slopes model, we found an SE of 25 ms, and a t-value of -6.6. Now we observe an SE of 16 ms and a t-value of -9.5. This corresponds to a downward bias of just over one-third, i.e., the biased version is 64% as wide as the original. In the case of smaller effect sizes, this difference in SE could lead to false conclusions about whether they are real or not, and consequent failures to replicate.

In the current example, the need to account for heterogeneity is bolstered by the size of heterogeneity (SD) in relation to the fixed effect (Mean), which is 140 ms/162 ms, which equals .86 units. This means that the predicted range of effects in the population are from 1 – 2(0.86) units to 1 + 2(0.86) units, which is from -0.72 to 2.72 units. Because the consequences of failures to model heterogeneity can be severe, it is important for researchers to understand how to determine the extent of the heterogeneity.

# Extent of Causal Heterogeneity: Noteworthy v. Ignorable

We have seen results where the extent of heterogeneity was noteworthy in the sense that it undermined the idea of a common causal process for the effect of stimulus valence on reaction time for self-descriptive words. In this section we will provide two examples, one where the heterogeneity is noteworthy and one where it is not. In doing so we provide guidelines as to when heterogeneity is sufficient to qualify conclusions of repeated-measures experiements.

## Noteworthy: Face-Orientation Effects

The following are data from a study by Sklar, Goldstein, Hassin and colleagues on the effects of spatial orientation on how fast faces emerge from suppression. The study had three facial orientation conditions, Upright, 90 Degrees, and Upside-Down. To simplify the presentation, we will focus on the Upright vs. Upside-Down conditions only. There were 21 participants and approximately 60 stimuli per condition, yeilding a total of 2544 observations. The results for the Upright vs. Upside-Down comparison are only minimally different from the equivalent results obtained in a three-condition analysis. In addition, because the data had a skewness value of 2.2 (which is greater than the common cutoff of +/- 2 for data to be considered non-normal), we used logRT as the DV.

The dataset and full SPSS and R syntax are available in the Supplemental Materials. Below are excerpted portions of syntax and output.

SPSS:

MIXED logrt WITH fo1

/FIXED=fo1| SSTYPE(3)

/METHOD=REML

/PRINT=TESTCOV G SOLUTION TESTCOV

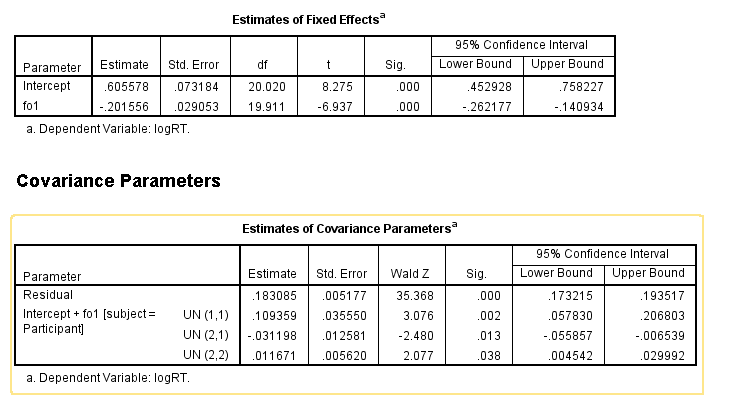
/RANDOM=INTERCEPT fo1 | SUBJECT(participant) COVTYPE(UN).

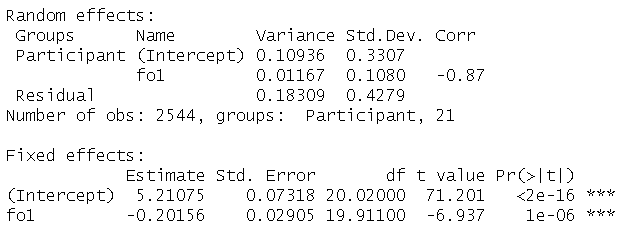
R:

ranfacem1 <- lmer(logRT ~ fo1 + (fo1 | Participant), data = ranfacefo1)

summary(ranfacem1)

confint(ranfacem1)

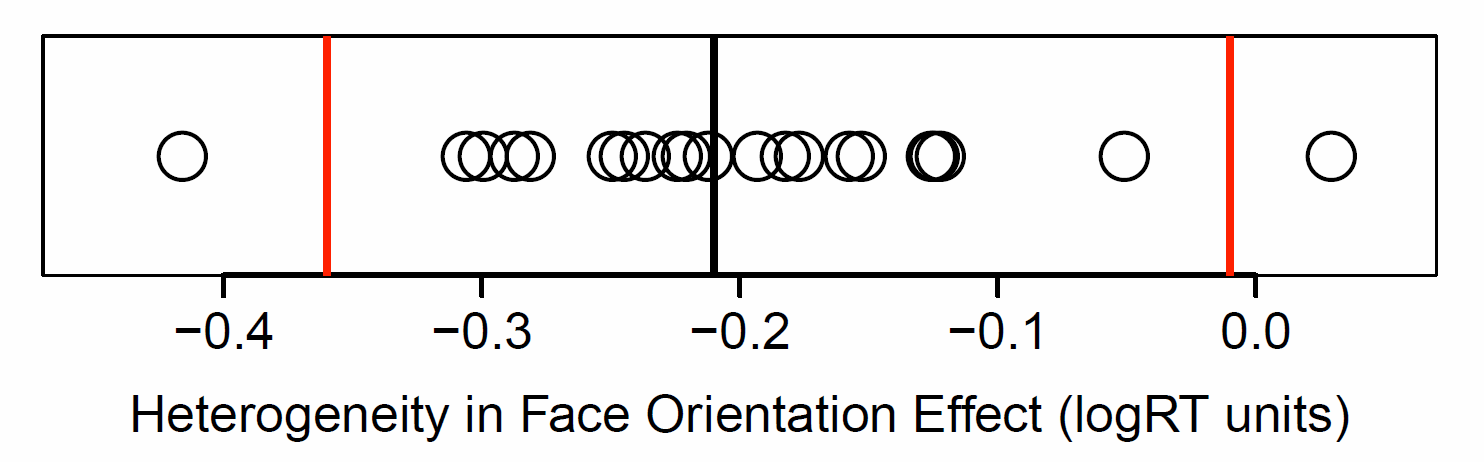


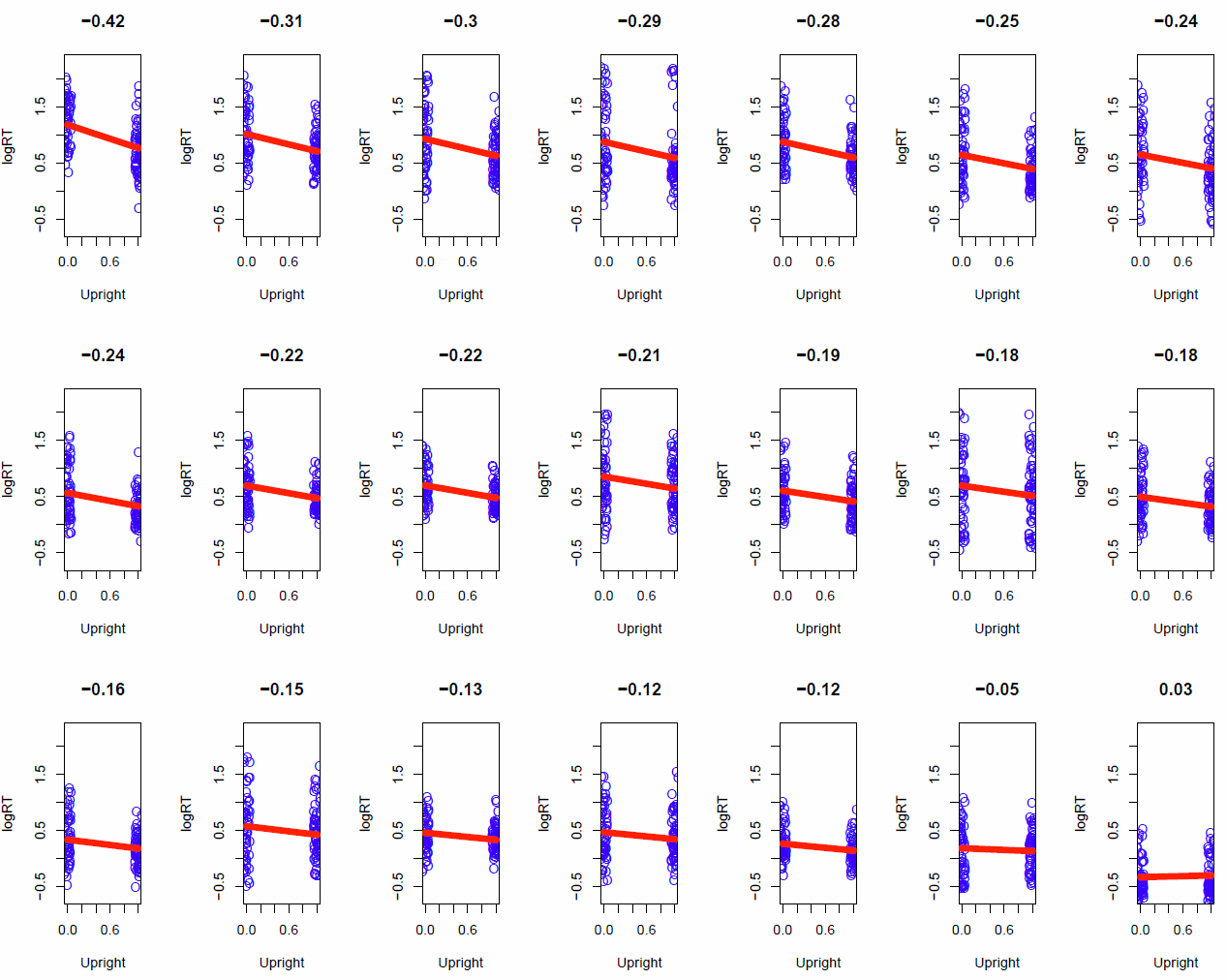


|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Effect | Population Estimates | |  | Population Heterogeneity | |
|  | Mean | SD |  | 2.5% | 97.5% |
| Level () | 5.21 | 0.33 |  | 4.5 | 5.9 |
| Orientation () | -0.20 | 0.11 |  | -0.43 | 0.03 |
|  |  |  |  |  |  |
| CI95 (Level) | [5.1, 5.4] | [0.24, 0.45] |  |  |  |
| CI95 (Orientation) | [-0.25, -0.14] | [0.06, 0.16] |  |  |  |

The average person is - .20 logRT units faster at responding to an upright versus and upside-down face (CI: [-0.26, -0.14]), with heteroegeneity of 0.11 SD units. This heterogeneity estimate is just over half the size of the fixed effect estimate, which we regard as substantial. Its size implies that the prediction interval for the face orientation effect ranges from 0.02 to 0.42 logRT units. This population prediction is mirrored in the range of values in the dotplot and panel plots above. A person at the 2.5th percentile of the distribution shows no face orientation effect, whereas a person at the 97.5th percentile shows an effect of xxx, twice as large as the effect for the average person.

How confident can we be that the model’s estimate of population variability, 0.11 SD unit, could not have occurred by chance? SPSS reports a confidence interval for the random slope of [0.07, .17]. The equivalent estimate in R (using the confint ( ) function) is [0.06, 0.16] in R. These estimates suggest that zero is a highly unlikely value. A second way to assess the importance of the heterogeneity effect is to compare the deviance scores of a model with a random slope for face orientation and one without. We will not concern ourselves with the details of this test (the details are in the supplementary materials) except to say that the Chi-squared test that results from the comparison suggests the addition of the random effect substantially improves the fit of the model to the data (Chi-squared(2) = 27.9, p = .009 x 10-4).





Thus heterogeneity can be clearly seen in the size of the random effect SD parameter, in the predictions for sample data, and in statistical indices that reject the notion that the observed effect could have occurred by chance. These details show, in our opinion, that any account of the experimental results that neglect the heterogeneity would be a marked distortion of the results, and would obscure important information for explaining the phenomena.

## Ignorable: Math Priming Effects

Although we think that causal heterogeneity is the rule in psychological and social processes, we acknowlege that in particular instances it may be sufficiently small to be ignorable. The next study provides an example of the latter instance. It is a dataset with results already reported in a PNAS paper by Sklar et al. (2012). In the original analyses using repeated-measures ANOVA, the question of heterogeneity was not addressed. Here we present the results using the mixed-model approach we have been recommending.

The study examines RT to pronounce simple numbers depending on whether subjects were subliminally primed with equations that yield this number (labeled “congruent”) or not (labeled “incongruent”). The orginal results indicated a substantial congruency effect, indicating that simple math operations are processed subliminally. For consistency with other studies in this paper, we report analyses using logRTs rather than raw RTs. The dataset and full SPSS and R syntax are available in the Supplemental Materials. Below are excerpted portions of syntax and output.

MIXED logrt WITH congruec ptimec

/FIXED=congruec ptimec congruec\*ptimec | SSTYPE(3)

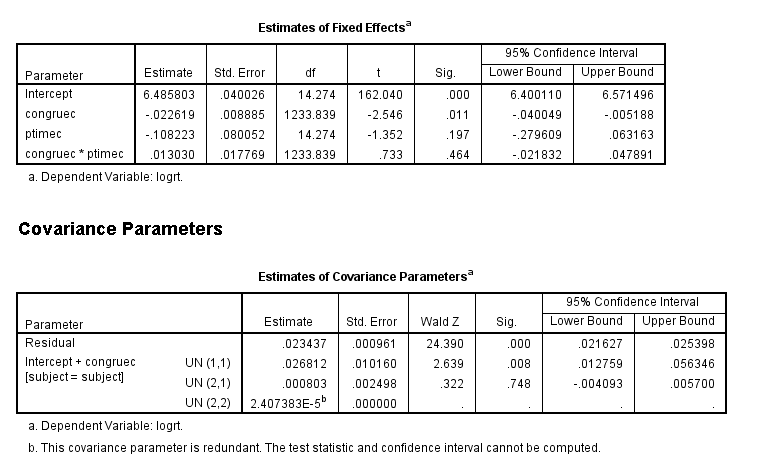
/METHOD=REML

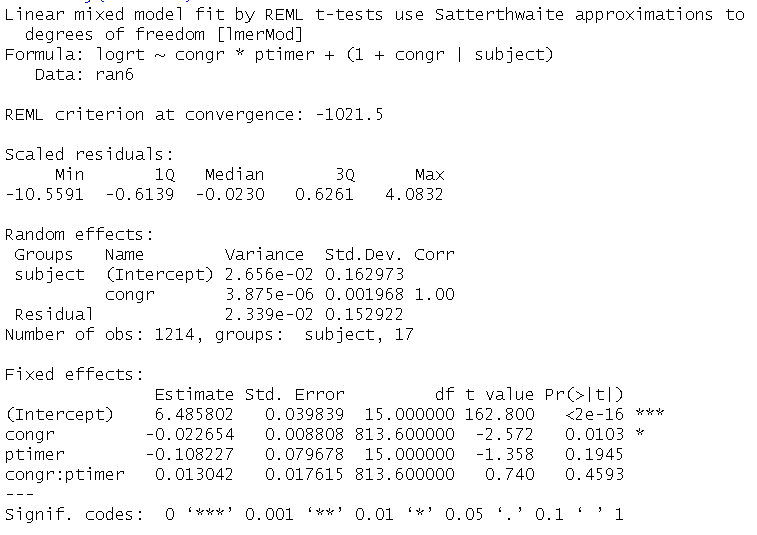
/PRINT=DESCRIPTIVES G SOLUTION TESTCOV

/RANDOM=INTERCEPT congruec| SUBJECT(subject) COVTYPE(UN).

ran6mod2 <- lmer(logrt ~ congr\*ptimer + (1 + congr | subject), data = ran6)

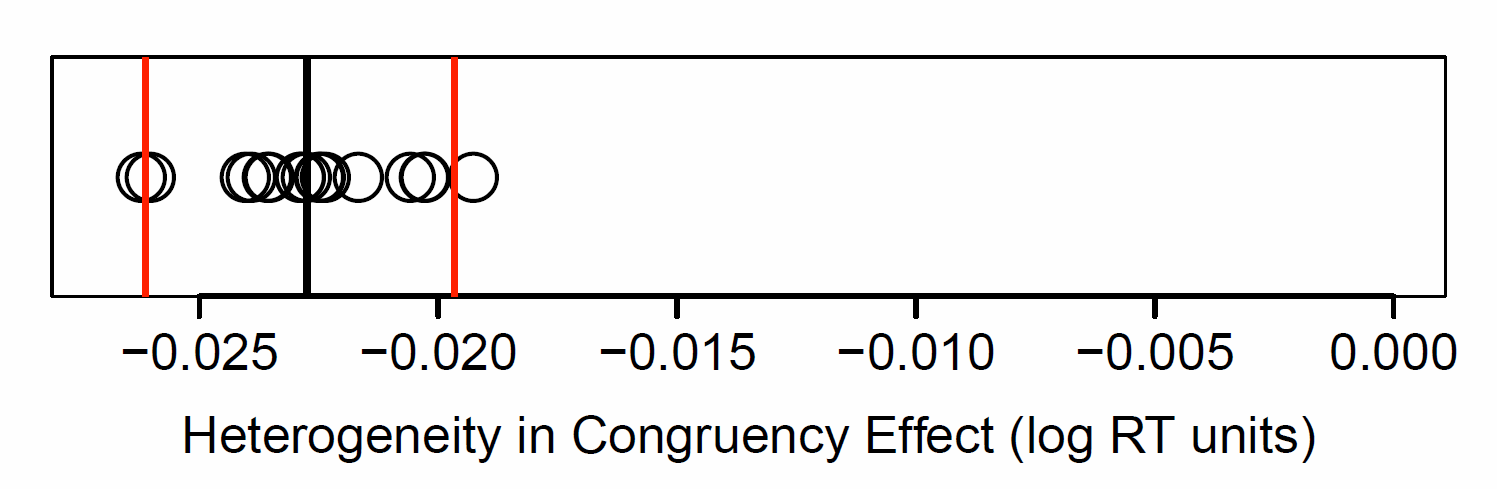
summary(ran6mod2)

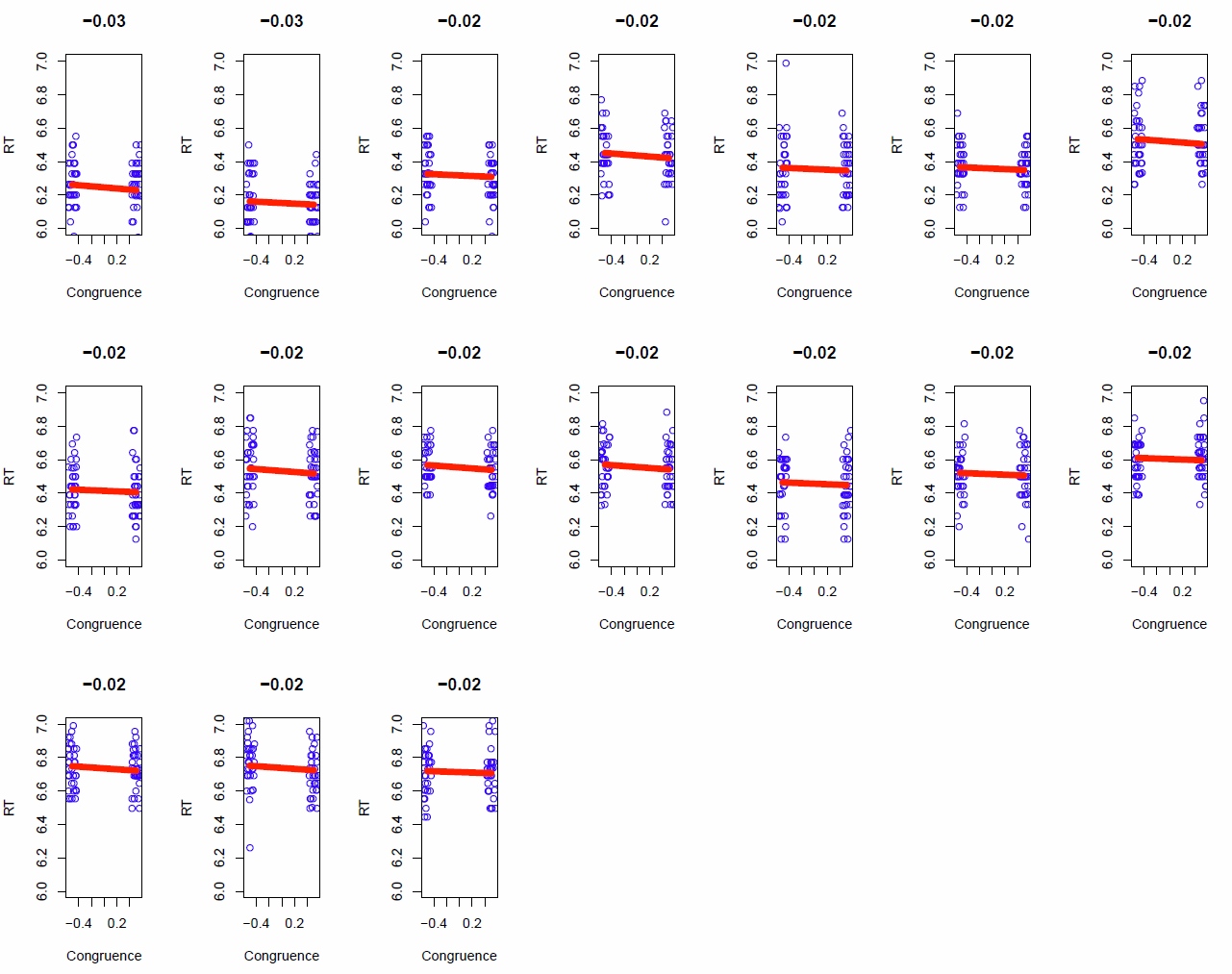




|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Effect | Population Estimates | |  | Population Heterogeneity | |
|  | Mean | SD |  | 2.5% | 97.5% |
| Level () | 6.49 | 0.16 |  | 4.5 | 5.9 |
| Congruence () | -0.023 | 0.002 |  | -0.019 | -0.027 |
|  |  |  |  |  |  |
| CI95 (Level) | [6.4, 6.6] | [0.11, 0.22] |  |  |  |
| CI95 (Congruence) | [-0.25, -0.14] | -- |  |  |  |

The results from SPSS and R indicate that there is a effect of congruence on logRT: -.023 units, CI [-0.05, -0.41]. SPSS Mixed reports that the heterogeneity parameter for congruence is redundant with heterogeneity in the intercept. R provides an estimate of -.002 SD units together with a standard error but reports a perfect correlation with the intercept variance. What this means is that allowing each person to have their own individual congruence effect gives us virtually no additional information beyond differences in their overall level of responding. The R results permit us to display predictions and plots, and although these are arguably of limited value, we report them because they give an indication of how small the estimated heterogeneity is. Based on the R results, the implied population range of effects for the congruence effect is narrow: from -0.019 to -0.027. The predictions for the sample, displayed in the figure above, are in a similar range. These indicate that the heterogeneity effect was less than one-tenth of the fixed effect value. The panel plots of these predictions show essentially no variation in slopes. Although R was unable to calculate confidence intervals for the congruence heterogeneity parameter, we were able to assess the contribution of heterogeneity to the fit of the model. The Chi-squared value of 0.044 with 2 degrees of freedom had a p value of .98, indicating the parameter’s minuscule contribution to model fit. In this case, we can confidently conclude that the priming effect is, in effect, the essentially the same across subjects.





## How to Decide if Heterogeneity Matters

In the two examples above, we used three criteria to decide whether the heterogeneity in an experimental phenomenon is sufficiently large to warrant attention. The first was whether the confidence interval for the heterogeneity parameter included (or was very close to) zero. This was possible for the face orientation data but not possible for the priming data (because of its minimal heterogeneity, given that variance effects cannot be negative). The second was whether the model fit was improved by allowing for heterogeneity. We saw clear evidence that it was for the face orientation data and that it was not for the priming data. The third was whether the size of the heterogeneity effect was large in relation to the effect for the average subject. For the face orientation data it was just over half the size; for the priming data, it was less than one-tenth the size.

For this last criterion, it is worth considering what level of heterogeneity is noteworthy. We suggest that heterogeneity is noteworthy if it is .25 SD units of the fixed effect or greater. Such heterogeneity implies that the 95% of the population have effect values that lie between 0.5 and I.5 times the effect for the average person. Thus an individual at the low bound of the distribution has an effect size that is half as great as that of the average person, and an individual at the high bound has an effect size that is 50% greater than that of the average person. Using this threshold, we conclude that the level of heterogeneity in the first two example datasets are noteworthy. In the face-orientation data, for example, the random effect is 0.5 SD unit of the fixed effect. This implies that the 95% of the population have values that lie between 0 and 2 times the fixed effect, a large range. In contrast, for the priming data, the prediction interval lies between .86 and 1.17 times the fixed effect. The dotplots presented for each set of results show this difference in heterogeneity very clearly.

# Explaining Causal Heterogeneity

As we have argued, discovering causal heterogeneity, even without knowing its sources, can be a contribution to understanding a phenomenon and a potent sign as to what subsequent experiments should focus on. Some researchers, however, might have theoretical predictions regarding those sources, and included suitable background measures of these in their experiment. In this case, the mixed-model analysis above can be expanded to include these measures. Even if such measures were not included in the study, some researchers may wish to conduct exporatory analyses using demographic or other variables available to them, or consider which theoretical models could help account for such between-person variability.

This section presents an example of the first case. Drawing on Regulatory Focus Theory (Higgins, Higgins & Higgins, 1980-2080), we examine the prediction that a promotion-focused motivational orientation, which emphasizes eagerly pursuing ideals and aspirations, will speed the endorsement of positively valenced traits whereas a prevention-focused motivational orientation, which emphasizes vigilantly pursuing duties and obligations, will speed the endorsement of negatively valenced traits. We draw on the same dataset presented in Study 1, but here we include promotion and prevention as moderating variables in our analysis.

We add to the original mixed-model, a between-subjects predictor, promotion orientation. The modified SPSS and R code are:

MIXED logrt WITH valenceE prom.c prev.c

/FIXED=valenceE prom.c prev.c valenceE\*prom.c valenceE\*prev.c prom.c\*prev.c

valenceE\*prom.c\*prev.c | SSTYPE(3)

/METHOD=REML

/PRINT=G SOLUTION TESTCOV

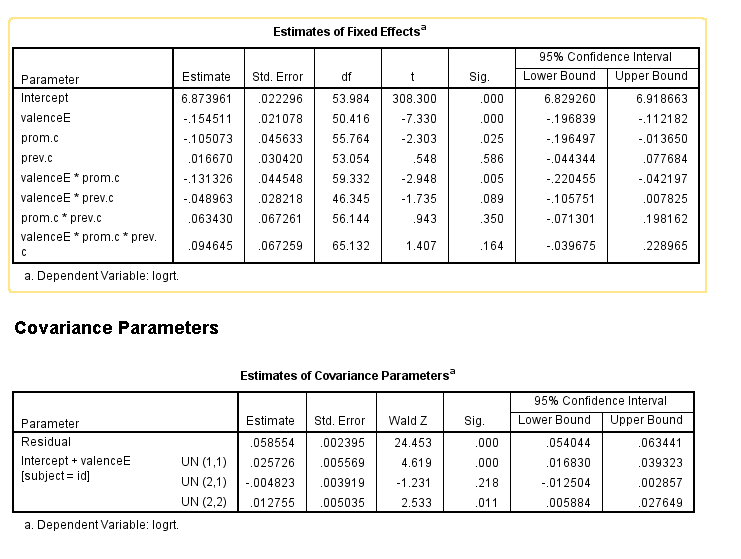
/RANDOM=INTERCEPT valenceE | SUBJECT(id) COVTYPE(UN).

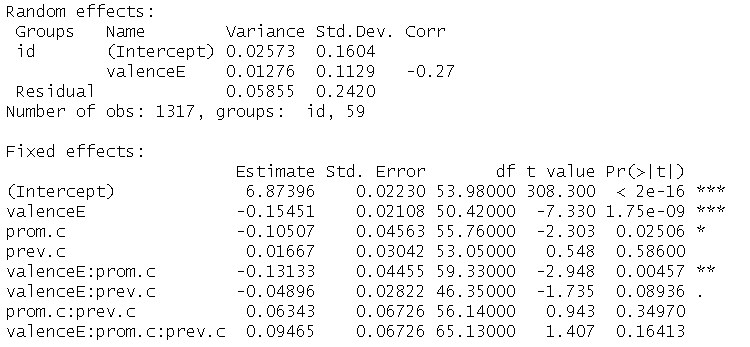
>modindiv1 <- lmer(logrt ~ valenceE\*prom.c\*prev.c + (1 + valenceE| id), data=datamerge)

>summary(modindiv1)

>confint(modindiv1)

The following are selections from the outputs of these programs:





Again we see no difference between the outputs. What we observe is evidence that both promotion and prevention predict the valence effect, but promotion is much stronger than prevention. For each SD unit difference in promotion, the model predicts that the speed advantage of positive traits differs by -0.13 logRT units. Thus the -0.15 logRT speed advantage of typical participants increased to -0.15 – 0.13 = -0.28 logRT advantage for those + 1SD on promotion. In other words, participants higher on promotion responded much faster to positive than negative traits compared to those lower on promotion. For prevention, the equivalent difference is -0.15 – 0.05 = - 0.20 logRT units. The effect plots below provide a visual display of each effect. Noteworthy is the differential effects of each motivation: Promotion operates primarily to increase the speed for positive traits, whereas prevention (marginally) operates primarily to slow the speed for negative traits.

Macintosh HD:Users:zeekatherine:KZNB:valenceXprom_figure.pdfMacintosh HD:Users:zeekatherine:KZNB:valenceXprev_figure.pdf

However, the residual variability of the valence effect remains substantial. In the basic model, it corresponded to an SD of -0.125 logRT ms. In the expanded model, the residual SD is -0.113 logRT. Clearly the lion’s share of the variation is not attributable to promotion and prevention.

# Temporal Stability in Causal Heterogeneity

Causal heterogeneity can be a function of the experimental procedure and of temporary states of participants as they undergo the procedure. As we have just shown, it can also be attributable to more stable aspects of participants such as their motivational orientation. If the heterogeneity is due in part to temporally stable variables, it follows that the heterogeneity itself should show some temporal stability. In this section, we will investigate this using a earlier version of the current experiment that involved two experimental sessions, one week apart. This study has already been presented in Higgser and Higgser (2005), but it was not used to investigate the question of stability. For the analyses we present, we confined ourselves to the subsample of 21 participants who provided experimental data on both occasions.

Here we will not take the reader through the analysis approach as it is identical to that used for earlier studies, but the code to conduct these analyses are in the supplemental materials for the paper. Note, that these analyses were carried out using the lmer package in R exclusively. The first figure we present is a plot of the predicted valence random effects for the sample of participants. We can see that both distributions show similar means and variances.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Effect | T1 Population Estimates | |  | T1 Population Heterogeneity | |
|  | Mean | SD |  | 2.5% | 97.5% |
| Level () | 7.05 | 0.19 |  | 6.7 | 7.4 |
| Valence () | -0.070 | 0.093 |  | -0.26 | 0.12 |
|  |  |  |  |  |  |
| CI95 (Level) | [7.0, 7.1] | [0.14, 0.23] |  |  |  |
| CI95 (Valence) | [-0.13, -0.01] | [0.08, 012] |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Effect | T2 Population Estimates | |  | T2 Population Heterogeneity | |
|  | Mean | SD |  | 2.5% | 97.5% |
| Level () | 7.00 | 0.22 |  | 6.5 | 7.5 |
| Valence () | -0.096 | 0.135 |  | -0.38 | 0.19 |
|  |  |  |  |  |  |
| CI95 (Level) | [6.9, 7.1] | [0.20, --] |  |  |  |
| CI95 (Valence) | [-0.17, -0.02] | [0.13, 0.14] |  |  |  |



A scatterplot of the Time 1 and Time 2 predictions demonstrates clearly that those participants who showed relatively large random effects at one time point tended to be the same participants who showed large random effects the other time point.



# Summary

We have completed our demonstration of how to use mixed-models to reveal the existence, extent and implications of causal heterogenity in repeated-measures experimental data. Our first substantive example concerned valence differences in the speed at which self-relevant adjectives were endorsed. We showed that the causal effect of stimulus valence on RT had striking heterogeneity (ranging from 2.4 times the size of the typical effect to more than half its size in the opposite direction) and that this heterogeneity was invisible in standard treatments of such data using repeated-measures ANOVA.

Our recommended analysis approach was a mixed model and we showed how it could be accomplished using SPSS and R. We showed that tests using mixed models (unlike repeated-measures ANOVA) require that the heterogeneity be explicitly modeled and that a failure to do so can lead to spurious results, failures to replicate, and impoverished understanding of the nature of the phenomenon being examined.

Our next two substantive examples were intended to distinguish cases where heterogeneity is fundamental to understanding the data and when it is not. We introduced three criteria, namely contribution to model fit, uncertainty interval, and size relative to the average effect. We paid particular attention to relative size, because assuming the other measures are positive, this is the strongest indicator of substantive relevance.

Although demonstrating heterogeneity can be a worthy end in itself, many researchers will also be interested in explaining any heterogeneity that is observed. To show how this can be accomplished, we returned to our first data example and showed that the heterogeneity in the valence effectwas attributable in part to relatively stable motivational orientations, namely promotion and prevention.

Heterogeneous effects can be due to temporary participant factors such as being hungry or having slept poorly, or implementation factors such as fluctuating lab temperature or equipment problems. They can also, however, be due to relatively stable aspects of participants such as their living conditions, their attitudes and values or their personality. We showed in a final study with two lab sessions a week apart that the heterogeneity showed marked stability. Thus in this example at least, the heterogeneity was not merely a fleeting effect of the participant’s state at the time of the experiment or unintended idiosyncrasies of the experimental procedure, but something relatively enduring about the participant’s reactions to the experimental manipulation.

# Discussion

We sought to demonstrate that data from repeated measures experiments are much richer than can be seen from traditional analyses. We showed that mixed modeling approaches, coupled with suitable graphical displays, provide an opportunity for experimentalists to discover hitherto-hidden diversity in their experimental effects. In this section we will consider the implications of this heterogeneity for theory and methods in experimental psychology.

## Opportunities for theory

Discovering heterogeneity in experimental effects can have important implications for theory. To the extent that one’s theory specifies a homogeneous causal effect, then observing heterogeneity is an indication of the need for further theory development. Heterogeneity that shows temporal stability indicates the existence of subpopulations. Perhaps the theory needs to account for subpopulations that differ in the causal process. If the heterogeneity is sufficiently stong, it can be the case that substantial proportions of the population show no experimental effect, or even reversals of direction in the effect (as was shown in the two studies of valence effects of RT. In the case of the valence experiment, the theoretical framework posited heterogeneity and measures of motivational orientation helped explain a portion of it. More generally, though, heterogeneity involving null effects or reversals could pose a challenge to the theoretical framework of the experimenter.

The presence or absence of heterogeneity has interesting implications for chains of causal effects. In experiments on chains of causation, manipulation of homogeneous links can be more efficient than manipulating heterogeneous links. This may be expecially true when the focus is on an outcome later in the chain.

Finally, it may be that there is no current explanation for the heterogeneity, and that an explanation will take years or decades to emerge. In this sense, observed heterogeity can be a placeholder for theoretically relevent explanatory variables.

## Opportunities for methods

Heterogeneity can indicate random idiosyncrasies in the experimental procedure: some testing sessions conducted in summer heat or winter cold. Some experimenters can put subjects at ease whereas others cannot. Some experimental procedures evoke heterogeneity. Knowing that this is the case can lead to revisions of procedure.

Heterogeneity can be used diagnostically to choose stimuli. If an investigator finds that participants vary greatly in their responses to a particular stimulus and are uniform in their response to a different stimulus, they could use this information in making choices between the two.

Heterogeneity can also be used to choose participants. If one can predict heterogeneity then one could preselect participants for whom an experimental effect is large, allowing sample sizes to be small—but, of course, limiting generalizability (see Shrout et al., 2017).

As we showed earlier in the paper, a failure to model heterogeneity can result in inflated test statistics and Type 1 errors and more worrying failures to replicate findings. An implication of this result is that power analyses should build in estimates of heterogeneity. In general true average effects that are heterogeneous will require larger samples to detect than similar average effects with less heterogeneity.

## Conclusions

To better understand causal processes, we believe that it is best to work from the assumption that one’s experimental effect, if present, is heterogeneous. The advent of mixed models for experimental data allow one to take account of any heterogeneity in one’s effect. If there is no heterogeneity, as we observed in the math priming dataset, these models can confirm that this too. We hope to have convinced the reader that causal heterogeneity is a likely feature of any realistic portrayal of psychological processes, and that this realism forms a better basis for applying psychology too.

# References

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. doi: https://doi.org/10.1016/j.jesp.2013.11.008

# Appendix 1: Pilot Study Details

Half of the personality traits were positive and half were negative. The items were selected on the basis of a pilot study. Initially, 156 words were selected from a list of 555 previously-identified personality traits (Anderson, 1968) for which it seemed plausible that there would be variance on endorsement of the item as self-descriptive, and that seemed relatively average on length and frequency in the english language. 72 Participants on M-Turk took part in the pilot study, 8 of whom were excluded for failing an attention check, and 13 for clicking on the page less than 150 times, which would have been required given that there were 150 traits on one page (all of these participants took less than 75 seconds to complete each page, suggesting their use of fuzzing to fill in the page automatically). The final N was 51.

Participants rated the desirability of the personality traits on 5-point scales from 1 (very undesirable) to 5 (very desirable) and also indicated whether they thought they possessed the traits by circling “me” or “not me.” Only traits with the percentage of “me” responses ranging from 25% to 75% were selected, so as to discard the traits that only a few or almost every student considered descriptive of themselves.  This left 71 traits. 5 content-confounded traits relevant to speed (e.g. “quick”, “indecisive”) were removed. The top 20 most desirable were selected as the “positive” traits (these had desire scores ranging from 3.59 - 4.69, e.g., “wise”, “courageous”), and the bottom 20 were selected as the “negative” traits (desire scores ranging from 1.29 - 2.18, e.g. “insecure”, “boring”). The two groups differed significantly on desirability (b = 2.41, SE = .09, t = 28.29, p < 0.001).  (Initially, the top 20 and bottom 20 differed significantly on length (top 20 were longer words), so several of the longest words from the top were replaced with shorter words that were just below the top 20 in desirability, and one of the shorter words from the bottom 20 was replaced with a longer word to balance out conditions.) The words were entered into the [English Lexicon Project](http://elexicon.wustl.edu/) to retrieve their length and log frequencies of use in the English language. They did not differ significantly with regards to length (b = .50, se = .63, t = 0.79, p = .44), nor log frequency (b = .54, se = .35, t = 1.52, p = .14). The remaining 5 “neutral” words were used as practice stimuli at the beginning of the task.