# Abstract

Title: Causal Heterogeneity: Opportunities and Pitfalls in Experimental Data

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 be 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 or process-irrelevant variability in experimental procedure or instrumentation, this approach makes sense. If, however, it captures heterogeneity in the causal process being investigated, then this approach leads to lost opportunities for (a) theory development, and (b) for methodological development for adequately testing theory. Furthermore, certain approaches to heterogeneity can result in overconfidence in findings or even spurious findings. In either case, the result is failures to replicate and distorted literatures.

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. In within-subjects designs, however, where causal inference involves comparing subjects to themselves in other conditions, such heterogeneity can be examined directly and can yeild important theoretical and methodological insights. By examined directly, we mean that its existence and strength can be established without necessarily knowing the sources. Our goal in writing this paper is to show how this can be accomplished.

## Example of Opportunities and Pitfalls: Effects of Stimulus Valence on RT Dataset

Our first example dataset comes from a replication of Higgins et al. Study 3. 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 an effect that can be readily confirmed. Our intent in this section is to first show using standard software an adequate analysis of the data.

### 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) followed by the Rosenberg self-esteem scale (Rosenberg, 1965), the Regulatory Mode Questionnaire (Kruglanski et al,. 2000), 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.

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 different 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.

Each personality trait appeared only once, in a random order. 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.

### Recommended analysis

To examine this hypothesis, we will primarily work with a statistical model where stimulus valence is the single predictor and log reaction time is the outcome. This simple model can be used to develop the main points regarding heterogeneity of causal effects. Later we will consider the need to add additional information about the temporal ordering of the stimuli and 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. Later, as models become more complex, it will be necessary to proceed with analyses in R alone.

If we examine in Table 1 a portion of the data matrix we can see that each log RT observation for each subject is given a separate data line. Each line contains columns for the subject’s unique indentifier (ID number), the exact stimulus word used (such as “confident” and “anxious”), the valence of each word (coded -0.5 if negative and 0.5 if positive) and the temporal order in which each word was presented to this subject (e.g., 1st, 15th, etc.). As is usual, stimulus ordering was randomized and each subject encountered a different order.

Our model specifies that in the population, a person’s log 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 replications (Kirk, Winer, Maxwell & Delaney). As we will see, when estimated using current mixed-model software, the output of this model is much more informative about the existence and size of causal heterogeneity than the output of a traditional repeated-measures ANOVA model.

The model can be written as follows:









Some readers will recognize Eq. 1, 2, 3 as a multilevel model expressed in Bryk and Raudenbush’s (2000) notation. In Eq. 1, the logRT observed for subject j (1, 2, …62) for stimulus valence i (-0.5, 0.5) is a function of an overall level for subject j, , and an effect of valence, , that is specific to each subject j. In this example, subjects with a greater value for B0j would tend to respond to traits more slowly than subjects with a lesser value for B0j. Similarly, subjects with a greater value for B1j would tend to respond faster to positive than negative traits, whereas those with a negative value for B1j would tend to respond faster to negative traits. In Eq. 2, between-subject variability in overall levels is composed of a level for the typcial subject, , and , each subject’s specific deviation (positive or negative) from that average. In Eq. 3, between-subject variability in the valence effect is composed of an effect for the typical subject, , and , each subject’s specific deviation (positive or negative) from that average. Finally, Eq. 4 is the same information presented in a single, mixed-model equation form. It is obtained by substituting Eqs. 2 and 3 into Eq. 1. We assume that the deviations of actual logRTs from their predicted values, the , are normally distributed with the same variance within valence condition. In addition, we assume that the subject-level random effects,  and  are normally distributed with mean zero, variances  and , and covariance . For a more elaborate introduction to these eqauations and their assumptions, see Bolger and Laurenceau (2013).

We will begin by estimating this model using mixed-model software in SPSS and R. The datasets are in the corresponding binary form. The syntax requred to estimate the model in SPSS is as follows:

MIXED logrt WITH valenceE

/FIXED=valenceE | SSTYPE(3)

/METHOD=REML

/PRINT=G SOLUTION TESTCOV

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

For R, this syntax is:

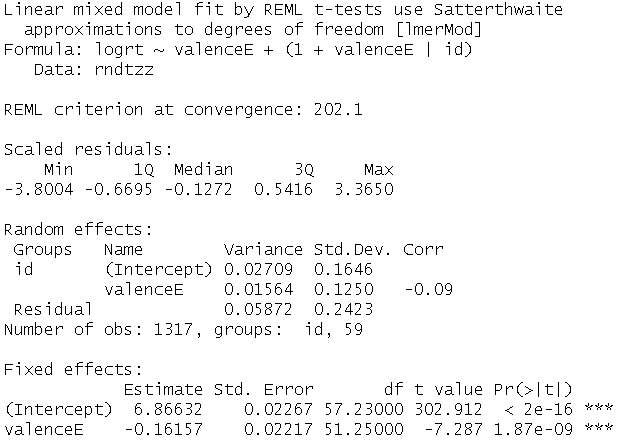
Out1 <- (lmer(logrt ~ valenceE + (1 + valenceE| id), data=study1a))

summary(Out1)

Both produce essential identical results.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Estimates of Fixed Effects** | | | | | | | |
| Parameter | Estimate | Std. Error | df | t | Sig. | 95% Confidence Interval | |
| Lower Bound | Upper Bound |
| Intercept | 6.866321 | .022668 | 57.228 | 302.912 | .000 | 6.820934 | 6.911709 |
| valenceE | -.161572 | .022173 | 51.249 | -7.287 | .000 | -.206080 | -.117064 |
| a. Dependent Variable: logrt. | | | | | | | |
|  | | | | | | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Estimates of Covariance Parameters** | | | | | | | |
| Parameter | | Estimate | Std. Error | Wald Z | Sig. | 95% Confidence Interval | |
| Lower Bound | Upper Bound |
| Residual | | .058719 | .002407 | 24.391 | .000 | .054185 | .063632 |
| Intercept + valenceE [subject = id] | UN (1,1) | .027087 | .005659 | 4.786 | .000 | .017985 | .040793 |
| UN (2,1) | -.001932 | .004034 | -.479 | .632 | -.009839 | .005974 |
| UN (2,2) | .015636 | .005594 | 2.795 | .005 | .007755 | .031527 |
| a. Dependent Variable: logrt. | | | | | | | |



It seems natural to begin by examining the estimate of the valence effect for the typical person, usually called a fixed effect in the mixed models literature (McCulloch, Searle & Niewhuis, 2012). The typical person is .16 logRT units faster at responding to positively valenced words than to negatively valenced words. The 95% CI is [-0.21, -0.12]. This difference corresponds to approximately 140 ms. For some researchers, this might be the only important statistic to report. In traditional repeated-measures ANOVA software, this is the result that is emphasized. But an average or fixed effect such as this is potentially a very limited and sometimes misleading estimate of a causal effect.

Before examining the other parameter estimates, we will turn to a graphical display of the raw data. Figure 1 is a panel plot showing each subject’s data for logRT as a function of valence. The panels are ordered by the size of the valence difference for each subject. The reader can see that the valence effect is discernable for many subjects, but it varies in size: The subject (id 35) in the upper left-hand corner shows a difference that is approximately [three times] larger than the subject in the middle of the panels (id 12). Heterogeneity such as this is part of the phenomenon to be explained, and the model proposed earlier will allow us to do so. To reiterate, although the fixed effect from our model revealed that the typical person was about 140 ms faster to endorse positively valenced traits as self-relevant compared to negatively valenced traits, it is apparent from this visualization that people differ in the magnitude and in some cases the direction of this effect.

First, the estimate of the SD of the valence random effect allows us to predict where 95% of the subjects in the population are. That is -.16 plus or minus 2\*0.125, which equals -0.41 to 0.09. This interval indicates that there are subjects at the 2.5% who have more than x times the average valence effect, and there are those at the 95% that have slightly reversed valence effects. This heterogeneity suggests that a one-size fits all view of valence effects that one gets from the fixed effect is quite misleading.

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 of Best Linear Unbiased Predictions (BLUPs). We display these in three ways. First we show them as a spaghetti plot in Figure x. The thick dark line is the fixed effect and the thin dark lines are the individual-specific effect for members of the sample. This illustrates the substantial causal heterogeneity observed: in this graph, the fixed effect line does not seem to describe the individual-specific effects lines very well. The second way we display predictions on the same page is as a stripchart, in which the estimated valence effect for each person is presented. These esimates are derived from extracting individual variability estimates from the random effects portion of the output and adding each individual’s estimate to the average effect. Again, the heterogeneity is evident. The third way is to overlay the predictions from our model on the observed paneled data in Figure x. Here we see that the model is capturing variation in the data quite well.

The reader will see the extra information contained in the results beyond the usual fixed effect estimates. For example, in this study, modeling the random effects gives us a idea of whether most people do indeed respond faster to positive than negative traits, or whether people substantially differ in the size of this effect. These results call for some theory that can explain these differences and measures that can predict them. In this case, the model suggests that there may be individual difference variables that could influence how quickly participants respond to positive or negative traits. We will consider this 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. The answer is that we do not have a balanced design. We are examining RT to traits that subjects say are self-descriptive and as can be seen from Figure x, fewer negative traits than positive traits are chosen by most subjects. Had we not selected on self-descriptive traits, we could 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 researchers who use repeated-measures ANOVA tend to aggregate their data before analyzing them. This not only results in incorrect inferences, but it also makes it impossible to estimate heterogeneity of effects and examine their properties. We now turn to a consideration of that approach.

### The aggregation approach

As we noted, 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.

The SPSS syntax to perform the aggregation is as follows: [have to figure this out]

In R it is: [copy relevant code]

To conduct a repeated-measures ANOVA on the aggregated data in SPSS, see the following input and output:

GLM rtneg rtpos

/WSFACTOR=valence 2

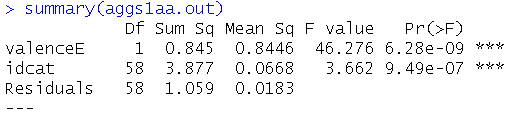
/METHOD=SSTYPE(3)

/WSDESIGN=valence.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tests of Within-Subjects Effects** | | | | | | |
| Measure: MEASURE\_1 | | | | | | |
| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. |
| valence | Sphericity Assumed | .845 | 1 | .845 | 46.276 | .000 |
| Greenhouse-Geisser | .845 | 1.000 | .845 | 46.276 | .000 |
| Huynh-Feldt | .845 | 1.000 | .845 | 46.276 | .000 |
| Lower-bound | .845 | 1.000 | .845 | 46.276 | .000 |
| Error(valence) | Sphericity Assumed | 1.059 | 58 | .018 |  |  |
| Greenhouse-Geisser | 1.059 | 58.000 | .018 |  |  |
| Huynh-Feldt | 1.059 | 58.000 | .018 |  |  |
| Lower-bound | 1.059 | 58.000 | .018 |  |  |

The R equivalent is:

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



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.

### 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. that was retracted from Psychological Science. In other words, although the model was specified correctly, it only allowed participants to differ in their average level of [whatever the DV was]. It did not include model terms that would allow researchers to examine how variable the effect of the IV was from person to person, hence causal heterogeneity was not accounted for in their original analysis. 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

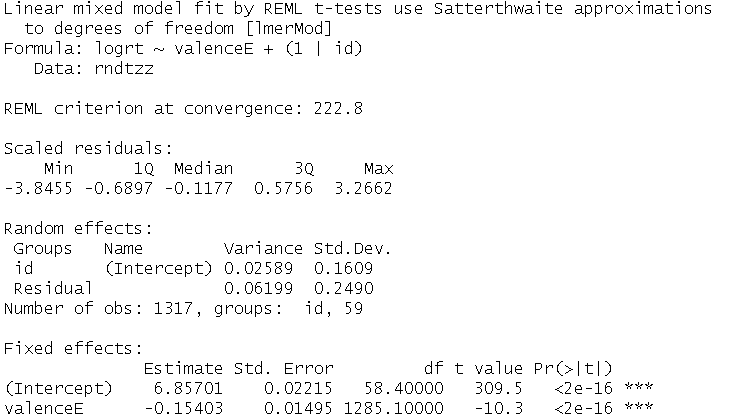
/RANDOM=INTERCEPT | SUBJECT(id) COVTYPE(UN).

R:

mle3aa <- (lmer(logrt ~ valenceE + (1 + valenceE| id), data=rndtzz))

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Estimates of Fixed Effectsa** | | | | | | | |
| Parameter | Estimate | Std. Error | df | t | Sig. | 95% Confidence Interval | |
| Lower Bound | Upper Bound |
| Intercept | 6.857012 | .022153 | 58.421 | 309.529 | .000 | 6.812674 | 6.901349 |
| valenceE | -.154034 | .014950 | 1285.127 | -10.303 | .000 | -.183364 | -.124705 |
| a. Dependent Variable: logrt. | | | | | | | |

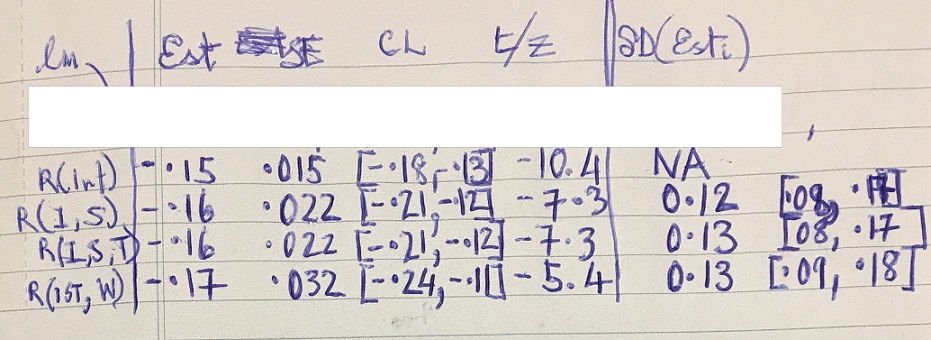
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Estimates of Covariance Parametersa** | | | | | | | |
| Parameter | | Estimate | Std. Error | Wald Z | Sig. | 95% Confidence Interval | |
| Lower Bound | Upper Bound |
| Residual | | .061994 | .002473 | 25.066 | .000 | .057331 | .067036 |
| Intercept [subject = id] | Variance | .025894 | .005356 | 4.834 | .000 | .017263 | .038839 |
| a. Dependent Variable: logrt. | | | | | | | |



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 0.022, and a t-value of 7.3. Now we observe an SE of .015 and a t-value of -10.3. This corresponds to a downward bias of almost one-third, i.e., the biased version is 66% as wide as the original. In the case of small effects, this difference in SE could lead to false conclusions about the exisitence of effects, and consequent failures to replicate.

### Summary

Summary table of the valence fixed (Est) and random (SD(Est)) effects for the various models. The first line of numbers are the effect estimates when the intercept is random but the valence effect is treated as zero. We see, as was shown earlier, the overly small standard error and the overly narrow CI. The second line are the estimates for the minimal model with both intercepts and slopes teated as random. The next line includes time as a fixed and random effect. This has no appreciable effect on the results. Finally, we have a model that also treats stimuli as random. In this example, certain stimuli may tend to be responded to more quickly or slowly. Here we see an increase in the SE for the fixed effect, but essentially no effect on the random effect variance of CI.



## Example of low heterogeneity dataset: Hassin et al., Experiment 7

## Example of high heterogeneity dataset: Hassin valence dataset

## Example of stability of random effects over time: Yuka dataset

## Benefits for Developing Theory and Methods

Heterogeneity can provide opportunities for theory development. For example, if a manipulation is inneffective for a third of participants in a sample, this suggests that covariates should be examined to determine if they help explain the variability in effectiveness. As far as opportunities for methodological development or improvement, heterogeneity can indicate systematic errors in experimental procedure. 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. A final

Causal chains: What links show relative homogeneity/heterogeneity? Where in the causal chain to intervene experimentally?

## Four Pitfalls

### Models that fail to account for heterogeneity at all

### Models that rely on averaging over replications. These can provide valid tests of fixed effects, but they cannot examine hetergeneity.

### Models that correctly take account of heterogeneity but almost always neglect to estimate the size of the effect.

### Models that correctly reveal the heterogeneity but this heterogeneity is not interpreted or sometimes not even discussed or reported.

# Discussion

## We saw that data from repeated measures experiments are much richer than can be seen from traditional analyses

## Opportunities for theory

### If there is heterogeneity, what does it tell us about the population being studied? Are there subpopulations for whom the process in distinct?

### Causal chains: What links show relative homogeneity/heterogeneity? Where in the causal chain to intervene experimentally?

## Opportunities for methods

### Can indicate systematic errors in 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.

## Problems and pitfalls

### Precision of estimates; power to detect random effects

### Difference between in-sample predictions and population estimates

## Conclusions: To better understand causal processes

### Best to assume heterogeneity than the opposite

### Tests are misleading: too liberal if they are ignored, too conservative if commonly methods of aggregation are used.

### A more realistic picture of psychological processes

### A better basis for applying psychology too

# Extra

## Sources of heterogeneity: Math priming

### Irrelevant to causal process

#### Experimental procedure: experimenter is unintentionally rude to participant. S’s chair causes discomfort.

#### Subjects: S who comes to lab is tired, hungry, in bad mood

### Relevant to causal process

#### Educational level

#### Math training

#### Attitudes toward math

## Non-Sources: Math priming

### Stimulus heterogeneity: specific math calculations within congruency condition

### Measurement error, other sources of random error at the trial level: standard the model separate this error from subject level heterogeneity

## Sources: Valence of self-descriptions

### Relevant to causal process

#### Promotion/prevention focus

#### Self-worth

#### Western/Eastern culture

## Non-Sources: Valence of self-descriptions

### Stimulus heterogeneity: specific self-descriptions