A Framework For Approaching Heterogeneity

B.A Edmar

A Framework for Approaching Heterogeneity

A key concepts within statistics is to analyse how variation in a value can be reduced by introducing explanatory variables which can lower the average squared distances from the mean. For example, using ANOVA we can try to reduce the variation in a dependent value by accounting for the variation in the effect introduced by the different treatments and within participant variation if we have repeated measures. A similar perspective can be employed within meta-analysis, but instead of using the term variance we use the word heterogeneity.

A common typology of heterogeneity is that from Cochrane institute which distinguished between clinical, methodological and statistical heterogeneity. Clinical heterogeneity is the variation in effects that is introduced by the participants within a study, the methodological heterogeneity is variation introduced by the study design, and the statistical heterogeneity is the statistically quantified variation in effects\results.

Weiss et al 2014 provides a similar but slightly different view of how to evaluate heterogeneity in the effect of programs. Like Cochrane, they argue that heterogeneity has tree sources: Treatment contrast, within participants moderators, and contextual moderators.

Treatment contrast refer to the strength of the measured effect. Consider that we have a completely valid concept X, but when we study X we can get either large X or small X dependent on the stimuli we use to elicit effects. Say we want to examine pain and we agree that this can be measured by putting your hand in cold water, the degree of coldness can be a source of difference in effects that then contributes to unexpected heterogeneity observed across a set of k studies examining this treatment. This is an interesting source of heterogeneity since the concept might be the same, but the true effect is not.

Within participants moderators refer to variation coming from individual participants. Say that we do our cold water experiment and have a single stimuli of water with a certain coldness, but our sample consists of two unaccounted for subgroups – one with a high pain tolerance and one with low pain tolerance. This difference within participants affects the strength of the effect, and thus is a systematic source of variation which we cannot attribute to normal sampling error – thus providing heterogeneity.

Contextual moderating factors are factors external to the participants and separate from the stimuli provided to elicit the effect. They cover contextual aspects of the ‘’study environment” that can cause potential differences in effects. For example, say that the coldness of water could be moderated by the size of the cup you place your hand into, and that our experiment uses multiple cups which vary in size. This could artificially introduce variation in the effect due to the context of the measurement and again provide heterogeneity.

It is important to note that the contexts of these perspectives on heterogeneity differs, one is from medicine and the other from political science and policy. However, they both have striking similarities. The main difference is that Cochrane places less weight on stimuli strength, which presumably would fall into methodological heterogeneity.

Cook, Campbell and Shadish makes the point that experiments, which are often used in psychology and medicine but cannot always be used in policy research, are highly local but have very general aspirations. That is, an experiment answers a very specific question about a specific group of people at a specific point in time. As such, there is a conflict between the aim of a scientific inquiry and that which is answerable through an experiment. However, if the construct we examine through our experiment is valid, we can make these generalizations more confidently.

Validity

Internal validity

Internal validity, or more accurately put, local molar causal validity, refers to whether an experiment capture a causal relationship between outcome and treatment. Bu going through the more granular term, we can get insight into what this means. The word causal in local molar causal validity emphasizes that internal validity is about causal inferences, not about other types of inference that social scientists make. The word local emphasizes that causal conclusions are limited to the context of the particular treatments, outcomes, times, settings, and persons studied. The word molar recognizes that experiments test treatments that are a complex package consisting of many components, all of which are tested as a whole within the treatment condition.

Exceptionally put by Cook, Campbell and shadish - ‘’internal validity is about whether a complex and inevitably multivariate treatment package caused a difference in some variable-as-it-was-measured within the particular setting, time frames, and kinds of units that were sampled in a study.” Below i list some threats to internal validity.

1. Ambiguous temporal precedence
2. Sample selection bias: could cause the creation of a confounding subgroup
3. History: Outside events occurring in parallel with the treatment of interest
4. Maturation: Participant/data changing during the treatment

5. Regression to the mean: Extreme scores will often be followed by less extreme scores on other variables.
6. Attrition: Loss of data could by systematic
7. Testing: any tests conducted during a treatment could be confused with a treatment effect.
8. Instrumentation: the nature of the measurement may change across the duration of the treatment. Think how political scales change.
9. Additive and interactive threats: The presence of multiple threats might ave additive or interactive effects which might exacerbate the inappropriateness of inferences based on internal validity.

A threat to the internal validity of a study could justify not viewing that study as comparable to a procedurally similar study that avoided that threat. The degree to which this can be done is subjective, but i does highlight a condition of comparability. Random assignment sorts out many of these threats.

Construct validity

In psychology, or soft sciences in general, empirical findings are only interesting to the extent that they can be defended as an approximation of a general construct. Any non-natural unit(and maybe even some natural) fall under this assertion. If we are cannot define the construct we want to measure, we cannot defend any measures taken to understand that construct. Congruence between what we aim to measure and the operations we undertake to measure that thing constitutes construct validity. When applied to concrete scales, the term operational validity can be used to answer the question - does the thing measure the thing we want to measure? this is of course highly relevant to topics without natural units. The Campbellian way of fostering construct validity is to

1. Make the person, setting, treatment and outcome constructs explicit.
2. Select instances which match the explicit definitions of the construct
3. Analyse the difference between the theoretical construct and the observed construct
4. Revise the explication of the outcome constructs.

This often breaks down as researchers select poor instances to match their explicit definitions of the construct resulting in a kind of slippage between the empirical and the theoretical. A very reductionist view of the problem of construct validity is that of the problem of categories. How do we conclude that something belongs to either category x or category y? This excerpt of Cook, Campbell and Shadish describes this well:

''And in part, it is because of the abstract nature of the entities with which social scientists typically work, such as violence, incentive, decision, plan, and intention. This renders largely irrelevant a theory of categorization that is widely used in some areas-the theory of natural kinds. This theory postulates that nature cuts things at the joint, and so we evolve names and shared understandings for the entities separated by joints. Thus we have separate words for a

tree's trunk and its branches, but no word for the bottom left section of a tree. Likewise, we have words for a twig and leaf, but no word for the entity formed by the bottom half of a twig and the attached top third of a leaf. There are many fewer "joints" (or equivalents thereof) in the social sciences - what would they be for intentions or aggression, for instance?"

This illustrates the silliness in holding fully positivistic epistemological/ontological positions. A theory of constructs must:

1. Emphasize operationalizing each construct in several ways within and across studies
2. Probe the pattern match between the multivariate characteristics of instances and the characteristics of the target construct
3. Acknowledge legitimate debate about the quality of that match given the socially constructed nature of both operation and constructs.

If these conditions are satisfied, a coherent theory of a construct can be constructed. However, there are also many threats to construct validity. The following list includes some of them:

1. Inadequate explication of constructs: May lead to incoherent operation and construct relations
2. Construct confounding: Some constructs are inseparable, this needs to be accounted for when measuring.
3. Mono-operation bias: All operationalization in isolation will underrepresents the target construct and include noise from other constructs not under investigation.
4. Mono-method bias: inadequate diversity in measurements - one could say methodological heterogeneity.
5. Confounding constructs with levels of constructs: Sometimes operationalizations does not capture all levels of a constructs, this complicates extrapolations.
6. Treatment sensitive factorial structure:
7. Reactive self-report changes: The motivation of participants to be in a treatment condition could change when conditions are assigned(do not disclose conditions to participants)
8. Reactivity to experimental situation: The participants perception of the experimental condition is part of the experimental treatment. That is, treatment in lab A is not the same as treatment in lab B, if participants have a systematic reaction of lab B, the construct validity is challenged.
9. Experimenter expectencies: Participants can be subject to desirability biases, these are a part of the treatment construct that is tested.
10. Novelty and disruption effects: Some treatments might be very disruptive or novel to the participants, this aspect of the treatment needs to be included in the treatment construct description.
11. Compensaory equalization: The compensation for participation is a part of the treatment and thus needs to be included in the treatment construct description.
12. Compensatory rivalry: If participants are aware that they are being treated, some rivalry across conditions needs to be apart of the treatment construct descriptions.

13. Resentful demoralization: Participants receiving a ''bad" treatment might respond more negatively than otherwise.
14. Treatment diffusion: In some cases, services from a treatment condition might extend to participants in other conditions, making construct descriptions of the conditions difficult.

Some aspects of these conditions are handled just through blinded randomized control trials, however, this is not always something that we can do. In these cases paying attention to these threats are important and necessary for making causal inferences.

External validity

External validity is th extent to which inferences holds to and across different settings, participants, persons and outcomes. The paper by Lucas (2003) does a great job illustrating issues with external validity and generalizability. The following list are threats to external validity identified by Cook, Campbell and Shadish.

1. Interaction of the Causal Relationship with Units: An effect found with certain kinds of units might not hold if other kinds of units had been studied.
2. Interaction of the Causal Relationship Over Treatment Variations: An effect found with on treatment variation might not hold with other variations of that treatment, or when that treatment is combined with other treatments, or when only part of that treatment is used.
3. Interaction of the Causal Relationship with Outcomes: An effect found on one kind of outcome observation may not hold if other outcome observations were used.
4. Interactions of the Causal Relationship with Settings: An effect found in one kind of setting may not hold if other kinds of settings were to be used.
5. Context-Dependent Mediation: An explanatory mediator of a causal relationship in one context may not mediate in another context.

These threats could possibly be treated as criteria for whether synthesis of evidence is reasonable. These are not the only things needed for synthesis, but they could be apart of a workflow for whether evidence synthesis is possible. One very interesting aspect of external validity is how it relates to random sampling. In the Campbellian view, random sampling is to external like random assignment is to internal validity.

Exercises for Your Understanding

1. **Is there such a thing as a ''True" effect? If so, what is it? Is it an exact parameter or a distribution of values?**

   I believe a true effect does exist in reality, but whether it is a fixed point/magnitude is hard to say. A true effect is a relation between two or more things. The effect can

sometimes be generative of a concept. I feel as if a safe place to start thinking about true effects is through John Locke, who states: ''That which produces any simple or complex idea, we denote by the name *cause*, and that which is produces, *effect''*. Through this lens, a true effect is simply the constant thing caused by something else. That is, a true effect is *the* thing that can be derived from some other thing. Put in context of an experimental paradigm, a true effect can be described as the true counterfactual difference between the worlds where something occurred and where that something didn't occur. This is obviously an impossible thing to observe empirically, but we can with the help of experimental design artificially construct a couterfactual.

In light of this definition of effect, I think we need to make the concession that it is a point estimate. The true differences between states is not variable. If that argument were to be made – that the true effect is distributed across some other meta/latent effect – we can argued that to infinity, there is always another underlying effect causing the distribution of a true effect. We need to bite the bullet with true effects and assume that they are fixed values that do not change.

2. **When are true effects causal and when are the not causal? How does this impact our research question?**

   All effects are products of a cause, but not all effects are causes of some other effect. They can be correlated with other effects but they do not have to be related in any other manner. For example, Y can cause the effects X and Z, but X does not have to cause Z. X does not have to be unrelated to Z either, there common cause – Y is what cements their dependency. For example, say we have two effects, a sound, and a physical sensation in our hands. The cause of both of these effects is me clapping my hands. This does not mean that the sound caused the sensation or that the sensation caused the sound, but they occur together and are therefore related. For our research into heterogeneity, we need to divide types of heterogeneity since they have different causes. What we might be interested in how the different heterogeneity effects are related to each other. For this we need to think about the laws of causality – temporal presidency, co-occurance, and logical consistency. Let's come back to this after we have a more consistent view of heterogeneity topology.

3. **Where does a true effect live ontological and epistemicaly?**

   I believe it resides within critical realism. We have to assume a consistent realism for effects to have a meaning. That is, it needs to exist in reality for us to systematically and reliably observe it. However, we must acknowledge that observation is flawed, and that there will be variation in interpretation of the value and meaning of empirical observation.

4. **What is the difference between an ''effect" and a ''true effect"?**

   For our purposes, there is no difference between the general term effect and ''true effect" since we already have recognized an ''effect" to mean the fixed counterfactual difference.

One might make the distinction that the ''true effect" is the contextualized version of *the* effect, meaning that the ''true effect" has clear mathematical properties depending on the data generative of the "observed effect" – which is separate from both the general effect and the ''true effect". Another way of illustrating this might be trough looking at the observed effect $\hat{\theta}_k$. The observed effect differs only from the true effect through random/sampling error. However, this might not be true for the general counterfactual effect, though i might be wrong on this. In the end, they are effectively the same for our intents. The Cochrane handbook provides a pretty nice description of what they refer to as **effect measures**. What they mean by this is statistical constructs that compare outcome data between two interventions groups, that is, an effect size or an observed effect. I think this definition of an observed effect is good and to the point. It also shares language with validity theory (construct validity).

5. **Some articles claim that methodological heterogeneity is crucial for theory development/testing/generalization and the progress of science. And some others (try to find examples) found that methodological heterogeneity led to increased uncertainty. How can these viewpoints be reconciled?**

   I think the meaning of uncertainty in this case is important to define. If uncertainty just IS heterogeneity then increased heterogeneity is a natural thing to expect when making a conceptual replication. That is the amount of unexplainable variation introduced through the replication. One aspect of Linden and Hönekopp covers this, they find that conceptual replications have larger standard deviations than close replications, meaning that the variation in effect size is larger than expected just from random error. On of the main ideas behind conceptual replication is to widen the space of potential explainable values of a parameter, meaning that through our acquisition of knowledge we gain a better ability to predict an outcome. However, if this generalization only causes us to make worse predictions then we have a serious issue. The more general a research question become, the more dependencies it requires for a accurate prediction, so when we generalize without trying to account for the generality, we will be doing a big mistake.

   I think the concept of incorporating our knowledge of the phenomenon in our predictions marries the idea that generality is good for theory but also bad for precision.

6. **Does methodological heterogeneity cause statistical heterogeneity? When?**

   I think it always will be some miniscule increase in statistical heterogeneity that will be caused by methodological heterogeneity unless the true effects observed are distinct but identical. Say the outcome of method A is identical to method B, can we be sure that they truly are distinct in their methodological effect? This might be logically true, but it is also a theoretical question. In other words, how can we be sure of their internal validity? As methodological heterogeneity is explored it introduces statistical heterogeneity by virtue of adding complexity to the model if not mathematically adjusted for. However, we can have interesting edge cases. Consider a situation where we observe effect Y from 5 procedurally identical studies, we then replicate effect Y from another 5 studies who

are different in measurement from the previous experiment but procedurally identical to each other. In the fist synthesis we get a tau of 5, in the second we get a tau of 4, but when we pool the results and adjust for the different methodological approach, we get a tau of 3. One possible explanation of such a scenario is that we have a poor understanding of the effect to the extent that instead of capturing effect Y, we capture effect Z. I think it might be best illustrated through a Q distribution. Image the same scenario of studying effect Y, but both our methods of measuring are skewed in relation to the chi^2(k-1) distribution, but when we pool our results, the variation of the data is captured by the new chi^2(k-1) distribution. I should test this through simulation.

7. **Can it be that methodological heterogeneous studies have low statistical heterogeneity?**

   I don't see a reason for why there might not be cases like this. It appears to be uncommon but I do not doubt that some examples exist.

8. **When can we distinguish between methodological and clinical heterogeneity?**

9. **Do meta-analyses try to restrict the methodological heterogeneity of the included study in order to potentially reduce statistical heterogeneity? Is this a good approach?**

   It depends on what ''restrict" means. If restrict is post-hoc changing the inclusion criteria to reduce statistical heterogeneity i do not think that is very smart. However, exclusion of some possibly influential studies as a sensitivity analysis is good. I do not think there is any harm in exploring the garden of forking paths so long as you can keep track of where you have been and how much you have explored.

10. **How much does null-hypothesis testing have to do with the problem of heterogeneity and failure to replicate?**

11. **Can informative hypothesis testing as an alternative to null-hypothesis testing be a remedy? E.g. by having more power?**

12. **Is it fair to think about heterogeneity as unaccounted for variation? Or perhaps unexpected variance?**

13. **Are the Cochrane categories(clinical, methodological, statistical) of heterogeneity correct? How does it relate to Weiss?**

14. **Why are random effects assumed to be normal?**