Why internal validity is not prior to external validity

Johannes Persson & Annika Wallin
Lund University, Sweden
Corresponding author: johannes.persson@fil.lu.se

[**Abstract:** We show that the common claim that internal validity should be understood as prior to external validity has, at least, three epistemologically problematic aspects: experimental artefacts, the implications of causal relations, and how the mechanism is measured. Each aspect demonstrates how important external validity is for the internal validity of the experimental result.]

1) Internal and external validity: perceived tension and claimed priority

Donald T. Campbell introduced the concepts internal and external validity in the 1950s. Originally designed for research related to personality and personality change, the use of this conceptual pair was soon extended to educational and social research. Since then it has spread to many more disciplines.

Without a doubt the concepts captures two features of research scientists are aware of in their daily practice. Researchers aim to make correct inferences both about that which is actually studied (internal validity), for instance in an experiment, and about what the results 'generalize to' (external validity). Whether or not the language of internal and external validity is used in their disciplines, the tension between these two kinds of inference is often experienced.

In addition, it is often claimed that one of the two is prior to the other. And the sense in which internal validity is often claimed to be prior to external validity is both temporal and epistemic, at least. For instance, Francisco Guala claims that:

"Problems of internal validity are chronologically and epistemically antecedent to problems of external validity: it does not make much sense to ask whether a result is valid outside the experimental circumstances unless we are confident that it does therein" (Guala, 2003, 1198).

The claim about temporal priority is that we first make inferences about the local environment under study before making inferences about the surrounding world. The claim about epistemic priority is that we come to know the local environment before we come to know the surrounding world.

In the following we problematize the relation between external and internal validity. Our claim is that the two types of validity are deeply intertwined. However, we are not going to attempt to argue for the full claim. We argue only in favour of the part of the claim that is in conflict with the idea behind the internal/external distinction. The argument is directed at showing that internal validity *understood as prior to external validity* has, at least, three epistemologically problematic aspects: experimental artefacts, the implications of causal relations, and how the mechanism is measured. We exemplify the problems associated with experimental artefacts and mechanism measurement by cases from experimental psychology. Each aspect demonstrates how important external validity is for the internal validity of the experimental result.

We end the paper by presenting a different kind of test. Lee Cronbach claims that internal validity, as interpreted by the later Campbell, is a rather meaningless feature of scientific results. If we are right, a Cronbachian attack on internal validity in general must also be mistaken. Since on our understanding internal and external validity are intertwined a successful attack on internal validity would threaten to have adverse effects on external validity. To be consistent with our standpoint the particular conception Cronbach attacks should pinpoint other features than the concept of internal validity has traditionally been assumed to capture.

2) What is internal and external validity?

It is impossible to evaluate whether the perceived tension and the claimed priority of internal validity are justified unless we know more precisely what it is that we make internally valid inferences about and what this validity is supposed to consist in. Below we present three formulations of internal and external validity:

Campbell's early conception: "First, and as a basic minimum, is what can be called *internal validity*: did in fact the experimental stimulus make some significant difference in

this specific instance? The second criterion is that of *external validity*, *representativeness*, or *generalizability*: to what populations, settings, and variables can this effect be generalized?" (Campbell 1957, 297).

Guala's recent conception: "*Internal* validity is achieved when the structure and behavior of a laboratory system (its main causal factors, the ways they interact, and the phenomena they bring about) have been properly understood by the experimenter. For example: the result of an experiment E is internally valid if the experimenter attributes the production of an effect B to a factor (or set of factors) A, and A really is the (or a) cause of B in E. Furthermore, it is externally valid if A causes B not only in E, but also in a set of other circumstances of interest, F, G, H, etc." (Guala 2003, 1198).

Campbell's later conception: "In the new contrast, external […] validity involve[s] theory. Local molar causal validity [, i.e. internal validity,] does not. While this contrast is weakened in the principle of proximal similarity [i.e. external validity], I still want to retain it. The principle of proximal similarity is normally (and it should be) implemented on the basis of expert intuition. […] Our intuitive expectations about what dimensions are relevant are theory-like, even if they are not formally theoretical. Moreover, clinical experience, prior experimental results, and formal theory are very appropriate guides for efforts to make the exploration of the bounds of generalizability more systematic." (Campbell 1986, 76)

Campbell's early conception and Guala's conception show similarity in how they understand external validity. It is about how to generalize what has been found internally. Campbell's later conception differs from both in that the connection between local causal claims and general claims is weakened. The word "local" emphasizes that the claimed validity is limited to "the context of particular treatments, outcomes, times, settings, and persons studied" (Shadish et al. 2002, 54). Local causal claims are "molar" as well. Campbell exemplifies it in the following way: "For the applied scientist, local molar causal validity is a first crucial issue and the starting point for the other validity questions. For example, did this complex treatment package make a real difference in this unique application at this particular place and time?" (Campbell 1986, 69). There is no guarantee that molar claims refer directly to a potential cause. A true molar claim entails merely that

something in the complex it captures is a cause. The difference between Campbell's later conception and Guala's conception is considerable in that respect. Guala's internal validity requires that we understand the causal mechanism that operates in the local case. The later Campbell explicitly opposes such a view as generally true of internal validity. Applied scientists also need internal validity, but they can normally not analyse causation with such precision; "to stay with our problems, we must use techniques that, while improving the validity of our research, nonetheless provide less clarity of causal inference than would a retreat to narrowly specified variables under laboratory control" (Campbell 1986, 70-71). The difference between Campbell's earlier and later understanding of internal validity seems to be one of emphasis primarily. However, the difference between their views of external validity is more significant. External validity is not in general established through representative sampling, and it is not a matter of simple inductive generalisation. First, a cause has to be extracted from the molar situation and then the causal relation is exported to proximally similar cases.

For each of these conceptions there are epistemologically problematic aspects of internal validity. We will focus on three: experimental artefacts, the implications of causal relations, and the measurement of mechanism.

3) Epistemology—the problem of experimental artefacts

Can there be such a thing as an internally valid inference? That clearly depends on whether the methods we use guarantee that we see clearly, i.e. that what we see in the local environment is not in fact an artefact of something else. But some well-known "internally valid" results have in fact been generated by, for instance, the method of randomization or measurement used.

3a) Overconfidence—experimental artefacts

Overconfidence is a psychological phenomenon that refers to an overrating of the correctness of one's judgements. Typically, participants are asked knowledge questions such as "Which city has more inhabitants? Hyderabad or Islamabad?" and are asked to rate how confident they are that their answer on this particular question is correct on a scale

from 50% to 100%. Overconfidence occurs when the mean subjective probability assigned to how correct responses are is higher than the proportion of correct answers. In contrast a participant is calibrated if: "…over the long run, for all propositions assigned a given probability, the proportion that is true equals the probability assigned" (Lichtenstein, Fischhoff and Philips, 1982).

The overconfidence effect can, however, be made to disappear under certain experimental conditions. Some authors (e.g., Gigerenzer, Hoffrage and Kleinbölting, 1991; Juslin, 1994) have claimed that the overconfidence effect is simply an effect of unrepresentative sampling. The basic idea behind the critique is that participants need a certain amount of information in order to make a correct estimate of their performance on a task. When this is not available, they will instead draw on their more general knowledge of the area. If I have no clear intuition on whether Islamabad or Hyderabad is the biggest city in the question above, I might use the knowledge I have of my general competency in geography or what I know about the capitals of Asian countries to produce a confidence judgement. That means that if the knowledge questions are sampled in a skewed way so that they contain more difficult questions than are normally encountered, participants will exhibit overconfidence (i.e. miscalibration). If the knowledge questions posed are instead randomly sampled from representative environments, the overconfidence effect disappears (Gigerenzer et al., 1991; Juslin, 1994).

The early experiments investigating overconfidence were clearly internally valid in the sense that results were robust: The experimental stimuli produced judgments that had the properties of overconfidence. However, they appear to be experimental artefacts, and slight variations in the experimental set up will change the results. There are, however, even more serious allegations against overconfidence – allegations that are especially interesting in this context. In a second set of critique against overconfidence authors such as Ido Erev (Erev, Wallsten and Budescu, 1994) and Peter Juslin (Juslin, Winman and Olson, 2000) claim that overconfidence (and the related hard-easy effect which we will not discuss here) is a product of regression towards the mean. Overconfidence occurs because a participant responding to a difficult task (as the one described above) is more likely to overestimate correctness than underestimating it. In the extreme, a participant that responds at a chance

level cannot be underconfident given the scale 50% to 100% certain that the response is correct. This explains also why the representatively sampled knowledge questions (of intermediate difficulty) made the overconfidence effect disappear. The artefact is not produced by the knowledge questions as such, but depend rather on features inherent in the experimental situation: it is difficult to conceptualize a scale measuring certainty that would not have endpoints such as these.

4) Epistemology—the problem of causation

Whether there can be an internally valid inference also depends on the nature of what is inferred to. Normally, as we have seen in 2) the inference is causal. Now, there are many concepts of causation. Some of these are clearly of a kind that does not support inferences that are primarily internally. For instance, someone operating with a notion of causation similar to one of those that Kant, Hume, or Mill relied on will judge internally valid inferences to causal matters impossible. For each of those causal concepts the implications of causation, regardless of whether it has to do with the notion of sufficiency or necessity, go beyond the local environment. If there is a causal relation in the local environment it follows that this holds also outside this environment. And, trivially, it holds that if it does not hold outside the environment it cannot hold inside either. Hence such concepts of causation warrant neither the alleged temporal nor epistemological priority of internal validity.

It is in fact a long distance between traditional causal concepts and causation that is suitable for being primarily internally validly inferred to. However, more than one advocate of randomised controlled trials adopts a view on which an intervention study underwrites a positive causal inference. Consider the following quote from David Papineau:

"You take a sample of people with the disease. You divide them into two groups at random. You give one group the treatment, withhold it from the other [...] and judge on this basis whether the probability of recovery in the former group is higher. If it is, then T [treatment] must now cause R [recovery], for the randomization will have eliminated the

danger of any confounding factors which might be responsible for a spurious correlation." (Papineau 1994, 439)

This is excessively optimistic for reasons having to do with the possible artefacts of randomization (cf. Shadish et al., 2002, Ch. 2) and the more general points that we have already pressed, but that is, not the present point. Let us assume that randomization is successful in the desired respect. Papineau's modified position seems to rely on a concept of causation given which in the relevant cases causation is entailed by (i.e. is unproblematically inferable from) the fact that the relative frequency of R in the intervention group is higher than it is in the control. Thus, for instance, the concept of cause employed is not that causes are sufficient in the circumstances, nor that they are necessary. This is plainly not so since neither kind of causation is entailed by the experimental fact (cf. Persson 2009).

5) Epistemology—the measurement of mechanism

How mechanisms are measured has a strong impact on the results obtained. As we saw in the case of overconfidence the choice of measurements can have unintended side effects, but the relation between how stimuli are presented and the effects that are measured is more complex than so. An interesting example comes from psychophysics and concerns range effects, i.e., effects due to the fact that participants receive more than one experimental condition.

5a) Range effects– the measurement of mechanism

Poulton (1975) presents a number of different range effects demonstrating how the order in which stimuli is presented in itself affect the result, or the type of mechanism that is being observed (an "unbiased" perceptual judgment, or judgments mediated by range effects – in themselves mechanisms). We will use the simplest example, where the range in which a stimulus is presented influences how far apart different stimuli are judged to be. In the case of Figure 1 the slope of perceived distances between stimuli is radically different when the end points are L_1 and L_2, rather than S_1 and S_2 when $\varnothing$ represents the physical magnitude and ψ the subjective (perceived) magnitude.

[INSERT FIGURE 1 POULTON; SEE LAST PAGE]
Figure 1. Adapted from Poulton, 1968.

Since participants' pre-conceptions of what the range of stimuli is will affect their responses, the "external validity" of the stimuli (in this context how well the range it introduces, or the range the experimenter assumes, matches participants' pre-conceived range of stimuli) determines whether the results obtained *in the laboratory* correctly capture the features of the mechanism operating there. Hence, in cases like these, external validity is a requirement for internal validity. Note that this potentially false estimate of the function has prefect internal validity. Given the range, the stimuli really do cause the response, and we have a fair grasp of what the mechanisms are.
Poulton himself, however, treats the results differently than we do: "All experimental data are not equally valuable. A theoretical model is unlikely to be better than the data which has shaped it. If data are of restricted validity as a result of unrepresentative sampling or the independent variables or of uncontrolled transfer effects, a model based upon the data is not likely to have great generality. This is the case however much data the model can fit, provided all the data has been generated using the same inadequate techniques of sampling or experimentation" (Poulton 1968, 1). We do not disagree with Poulton, but in contrast to him we emphasize that the core issue here is how internal validity is to be guaranteed unless range effects are properly understood. And this will happen only when extra-experimental factors (such as participants' pre conception of the range that is to be introduced) are properly understood. Thus we would like to maintain that the case of the perceptual mechanisms at the mercy of range effects internal and external validity cannot be treated as separate entities.

6. The difficulty of adapting systems

A straightforward extension of the above observations about the co-dependence of external and internal validity is to be found in Egon Brunswik's work on representativeness. What he adds to the discussion is a focus on the difficulties in observing an organism that adapts

to the circumstances in which it exists: "The concept inherent in functionalism that psychology is the science dealing with the adjustment of organisms to the environment in which they actually live suggest the need of testing any obtained stimulus-response relationship in such a way that the habitat of the individual, group, or species is represented with all of its variables, and that the specific values of these variables are kept in accordance with the frequencies in which they actually happen to be distributed." (Brunswik, 1944, 69).

Note, however, that here the focus is exclusively on the adaptive character of human cognition (in Brunswik's case the perceptual system). If the aim of an experiment in psychology is to understand the functioning of different psychological mechanisms (in the form of stimulus-response relations), then the quality of this finding is just as dependent on whether the psychological mechanism has been properly activated as it is on whether the results can be replicated. This is not only a question about how the result will generalize to other settings (external validity) – it is a question about whether a proper result has at all been generated (internal validity). Thus, for psychological mechanisms that can be assumed to have an adaptive character, external validity (or certain aspects of it) appears to be prior to internal validity: It is more important that an experiment measures what it aims to measure than that the result internally valid.

6a Is the study object human cognition or the environment?

Egon Brunswik is one of the psychologists that have most clearly advanced the idea that external validity has to be taken into account if we are to understand the human mind at all. In his own words: "psychology has forgotten that it is a science of organism-environment relationships, and has become a science of the organism" (Brunswik, 1957, 6). His remedy to this difficulty was the notion of representative design (Brunswik, 1955), and, in particular, his use of representative sampling while studying perceptual constants (Brunswik 1944).

In his 1944 study, Brunswik wanted to understand whether the retinal size of an object could be used to predict its actual size. In order to establish the relationship between retinal size and object size, participants were followed for several weeks and stopped at random

intervals. For whatever object they were looking at, at that point, retinal size, object size, and distance were measured. Since the objects taken into account were the objects actually attended to by participants in their daily environments, Brunswik could estimate the real-life predictive power of retinal size for object size. His conclusion was that the retinal size had some predictive power regardless of the distance to the object.

Note that Brunswik's method as described here is *only* a method for understanding the environment. In order to explain how participants judge the size of objects, it has to be combined with a demonstration that retinal size is used to predict object size. However, the controlled experiment that can be used to test this hypothesis will not help us understand how predictive retinal size is of object size. This requires a method such as Brunswik's.

Note also that the method of representative sampling is only possible in so far as the researcher *already* has a clear understanding of the cognitive process under investigation. Unless we have some idea of which aspects of the environment are accessed by the cognitive mechanism, methodological shortcuts such as representative sampling are not possible. Simply stated, we have to know what to measure in order to measure it, also when the measurement is done through random sampling. Campbell, of course, notes this problematic issue in the context of random sampling of *participants* (note the difference in emphasis). He points out that: "… the validity of generalizations to other persons, settings, and future (or past) times would be a function of the validity of the theory involved, plus the accuracy of the theory-relevant knowledge of the persons, settings, and future periods to which one wanted to generalize […]. This perspective has already moved us far from the widespread concept that one can solve generalizability problems by representative sampling from a universe specified in advance" (Campbell 1986, 71).

Also other methodologically inclined psychologists have reflected upon the co dependency of the environment and the agent. Often this is conceptualized as the difficulty of understanding whether what is being observed is a feature of the participant's internal processing or a feature of the task environment. Thus Ward Edwards (1971) observes that: "My own guess is that most successful models now available [in psychology] are successful exactly because of their success in describing tasks, not people …modelling tasks is different from modelling people, [we need] to hunt for tools for modelling tasks,

and to provide linkages between models of tasks and models of people". And this difficulty has it roots in precisely the difficulty of making controlled experiments that observe features of a cognitive system designed for adapting to the circumstances. Or in Campbell's own words: "Both criteria [external and internal validity] are obviously important although it turns out that they are to some extent incompatible, in that the controls required for internal validity often tend to jeopardize representativeness" (Campbell 1957, 297).

7) Cronbach's challenge

Let us now set the objections against the possibility of internally valid inferences aside. Let us grant that the problems of randomization, measurement and causation can be dissolved by appropriate adaptive measures. Even so the question whether internal validity should be given priority remains:

"I consider it pointless to speak of causes when all that can be validly meant by reference to a cause in a particular instance is that, on one trial of a partially specified manipulation t under conditions A, B, and C, along with other conditions not named, phenomenon P was observed. To introduce the word cause seems pointless. Campbell's writings make internal validity a property of trivial, past-tense, and local statements." (Cronbach 1982, 137)

Cronbach's point translates nicely to what we have argued here. To the extent that there is a variety of causation that can be fully examined in such a way that it underwrites a positive causal inference—for instance, by a randomized controlled trial—then that variety of causation is not very scientifically valuable. What should we do with these past tense, local statements concerning highly artificial experimental contexts? They seem trivial as scientific results. The only way this kind of trivial causal statements could prove useful is if they connect with more substantial ones. In other words, internal validity of this kind could have a value in relation to external validity as providing one of the instances externally valid claims have to be true about. Now, internal validity is not prior to external validity in any interesting sense. If anything, it seems secondary. It should be noted that Campbell (1986, 70) acknowledges this: "The theories and hunches used by those who put

the therapeutic package together must, of course, be regarded as corroborated, however tentatively, if there is an effect of local, molar validity in the expected direction". However, this relationship between internal and external validity is important. Cronbach's challenge might be reconstructed as a counter argument to our claim that internal and external validity are intertwined. It might be constructed as the view that internal validity is redundant. As we have seen our response is: 1) to the extent that the causation internal validity concerns is substantial, external validity is needed as part of the evidence; 2) to the extent that the causation is of a trivial form, this kind of causation might still be important as one of the instances that is needed to prove external validity. (There is, of course, a third possibility as well, that all genuine causation is local.)

8) Priorities reconsidered

However critical we have been of attempts to prioritize internal validity, there is a last argument that can be made in its favour, and it is elegantly (and fittingly) made by Campbell in the following passage: "If one is in a situation where either internal validity or representativeness must be sacrificed, which should it be? The answer is clear. Internal validity is the prior and indispensable consideration. The optimal design is, of course, one having both internal and external validity. Insofar as such settings are available, they should be exploited, without embarrassment from the apparent opportunistic warping of the content of studies by the availability of laboratory techniques. In this sense, a science is as opportunistic as a bacteria culture and grows only where growth is possible. One basic necessity for such growth is the machinery for selecting among alternative hypotheses, no matter how limited those hypotheses may have to be." (Campbell 1957, 310). Although we do not believe that internal and external validity can be treated separately – or even chosen between in the way suggested by Campbell – we fully agree that scientific research will have to take whatever routes are available.

References

Brunswik, E. (1944). Distal focussing of perception: size constancy in a representative sample of situations. *Psychological Monographs, 56*(1), Whole No.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review* 62(3): 193 - 217.

Brunswik, E. (1957). Scope and aspects of the cognitive problem. In J.S. Bruner, E. Brunswik, L. Festinger, F. Heider, K. F. Muenzinger, C. E. Osgood and D. Rappaport (eds.). *Contemporary approaches to cognition*. Cambridge: Harvard University Press.

Campbell, D., T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54 (4): 297-312.

Campbell, D., T. (1986). Relabeling internal and external validity for applied social sciences. In W., M., K. Trochim (ed.). *Advances in Quasi-Experimental Design and Analysis. New Directions for Program Evaluation*, no 31. San Francisco: Jossey-Bass, Fall 1986.

Cronbach, L., J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass Publishers.

Edwards, W. (1971). Bayesian and regression models of human information processing – A myopic perspective. *Organizational Behavior and Human Performance*, 6: 639-648.

Gigerenzer, G., Hoffrage, U. and Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review* 98(4): 506-528.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science*, 70(5): 1195-1205.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of items. *Organizational Behavior and Human Decision Processes* 57: 226-246.

Juslin, P., Winman, A., and Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: a critical examination of the hard-easy effect. *Psychological Review* 107(2): 384-396.

Lichtenstein, S., Fischhoff, B. and Philips, L., D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman and A. Tversky (eds.). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

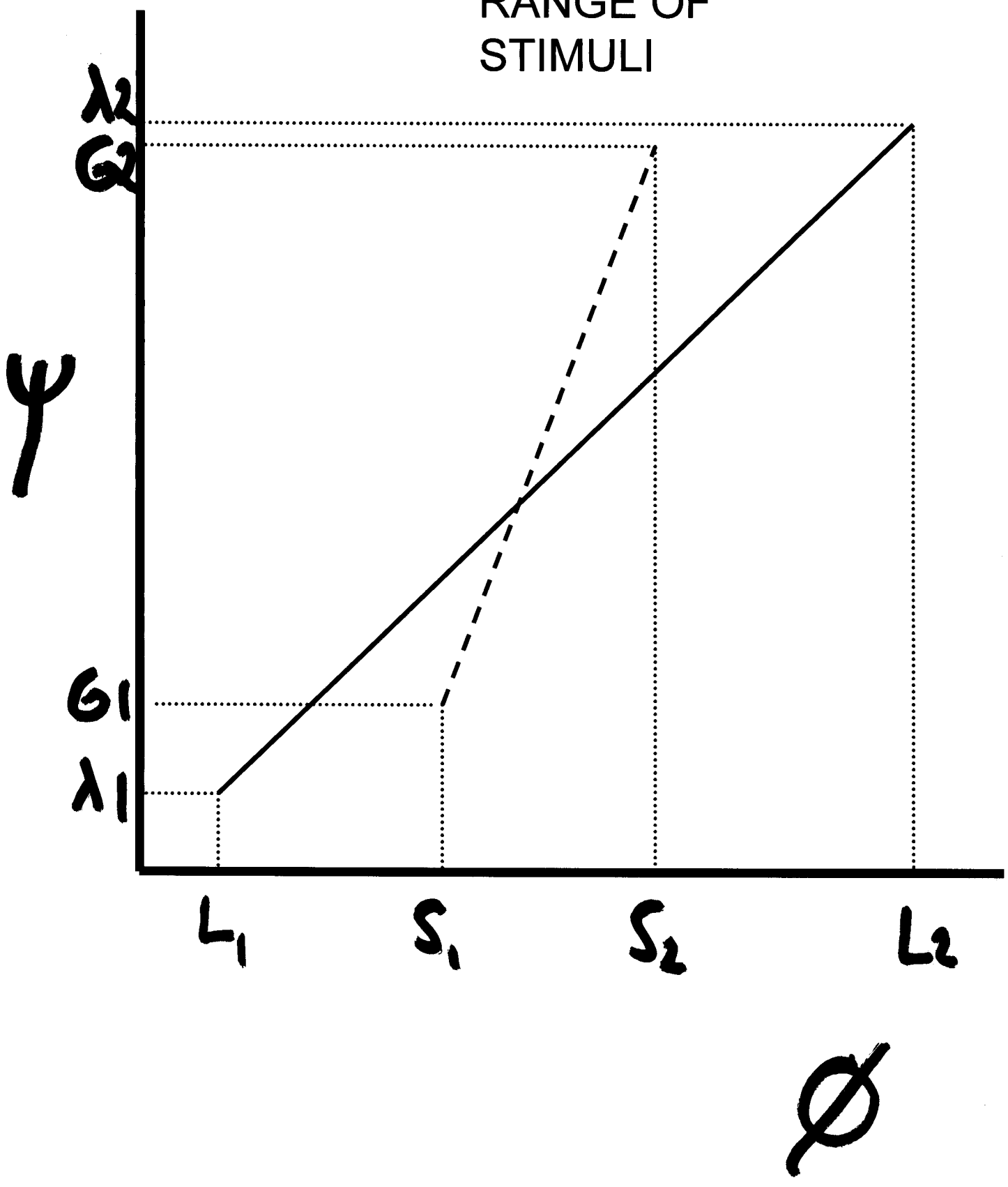
Papineau, D. (1994). The virtues of randomization. *British Journal for the Philosophy of Science*, 45(2), 437–450.

Persson, J. (2009). Semmelweis's methodology from the modern stand-point: intervention studies and causal ontology. *Studies in History and Philosophy of Biological and Biomedical Sciences* 40: 204–209

Poulton, E., C. (1968). The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin*, 69: 1-19.

Poulton, E., C. (1975). Range effects in experiments on people. *The American Journal of Psychology*, 88(1): 3-32.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasiexperimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.


RANGE OF
STIMULI
λ2
G2
ψ
G1
λ1
L1
S1
S2
L2
ϕ