The (misconceived) distinction between internal and external validity

Persson, Johannes; Wallin, Annika

Published in:
Against boredom: 17 essays on ignorance, values, creativity, metaphysics, decision-making, truth, preference, art, processes, Ramsey, ethics, rationality, validity, human ills, science, and eternal life to Nils-Eric Sahlin on the occasion of his 60th birthday

2015

Against boredom

FRI TANKE

Against boredom

17 essays

ON IGNORANCE

VALUES

CREATIVITY

METAPHYSICS

DECISION-MAKING

TRUTH

PREFERENCE

ART

PROCESSES

RAMSEY

ETHICS

RATIONALITY

VALIDITY

HUMAN ILLS

SCIENCE

AND ETERNAL LIFE

TO NILS-ERIC SAHLIN

ON THE OCCASION

OF HIS 60TH BIRTHDAY

EDITED BY

JOHANNES PERSSON

GÖRAN HERMERÉN

AND EVA SJÖSTRAND

The (misconceived) distinction between internal and external validity

JOHANNES PERSSON
ANNIKA WALLIN

1. Two common (but misconceived) claims about internal validity: the priority and trade-off claim

Researchers often aim to make correct inferences both about that which is actually studied (internal validity) and about what the results generalize to (external validity). The language of internal and external validity is not used by everyone, but many of us would agree that *intuitively* the distinction makes a lot of sense.

Two claims are commonly made with respect to internal and external validity. The first is that internal validity is prior to external validity since there is nothing to generalize if the findings obtained in, for instance, the experimental setting do not hold. The first claim is explicit in many writings. See for instance Francisco Guala's influential book *The methodology of experimental economics* (2005). And it is often implicitly relied on. The second claim is that researchers have to make a trade-off between internal and external validity. When one is increased, the other will decrease. The second claim was made already from the start by D.T Campbell in his classic *Factors relevant to the validity of experiments in social settings* (e.g., Campbell 1957, 297).

There is a certain tension between the first and the second claim. It has been argued before that it might be difficult to combine them. We intend to make the stronger point that both claims are misconstrued. Our hypothesis is that the relationship between internal and external

validity has to be re-conceptualized, and we will briefly indicate how.

2. *Some remarks about the origin of the divide between external and internal validity*

Donald T. Campbell introduced the concepts internal and external validity in the 1950s. In this text we rely on his 1957 classic (already mentioned in the introduction) as the primary source to his conceptual pair:

> 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)

The original article discussed research related to personality and personality change, but the conceptual pair of external and internal validity was soon extended to educational and social research. Since then it has spread to many more disciplines. Without a doubt the concepts capture – roughly, at least – two features of research that 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, researchers often experience the difference and sometimes the tension between these two kinds of inference. For instance, Nancy Cartwright in her *Hunting causes and using them* (2007, 220) calls the trade off between the two kinds of validity "a well–known methodological truism".

It is interesting to note that there in Campbell (1957) is no explicit mentioning of causal inference. On the other hand the language of effects is used rather extensively – as, for instance, in the above introduction of internal and external validity. What is salient already from the beginning is a strong link between the internal/external validity

distinction and the process of finding hypotheses among which a choice can be made:

> 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)

The causal vocabulary in Campbell's writings becomes more pronounced in his later production. At the same time, Campbell weakened his claims concerning the connection between local and general causal claims. There is a clear difference between Campbell 1957 and his *Relabeling internal and external validity for applied social sciences* from 1986, for instance. Partly, we think, this was because of his growing interest in applied sciences. Applied scientists also need internal validity, but they can normally not analyse causation with precision. There is a certain vagueness in the context of application. It is normally impossible to say with certainty which components of an intervention are causally relevant. This has implications for the internal/external validity distinction. At any rate this appears to be the received wisdom today, and it is reproduced in influential textbooks – such as in the *Experimental and quasiexperimental designs for generalized causal inference* written by W.R. Shadish, T. D. Cook and Campbell himself (2002).

3. On the priority claim: temporal and epistemic aspects

In both introductory and more advanced methodological textbooks, it is often claimed that internal validity is both temporally and epistemically prior to external validity. An example is Francisco Guala's paper *Experimental localism and external validity*:

> 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 (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. Maria Jimenez-Buedo and Luis Miller (2010) have recently collected a number of similar claims from the literature. Two examples are: "internal validity is a necessary but not sufficient condition for external validity" (found in *The challenge of representativeness design in psychology and economics* by Hogarth 2005); and "if there are doubts or questions about whether a relationship is real or spurious, then whether or not the finding applies to other settings is irrelevant" (found in *Reliability in experimental sociology* by Thye 2000).

The rising interest in experiments and methodological issues in sciences where experimentation has not been extensively used before has pushed the internal and external validity distinction into focus, although comparatively little – indeed, surprisingly little – has been written about the topic within philosophy of science. Recently, it has mostly been addressed in the philosophy of economics, due to the rising importance of, and philosophical interest in, experimental economics.

4. *The curse of context*

The discussion within philosophy of economics and philosophy of natural sciences interconnect. For instance, it is claimed by Jones (2011) in *External validity and libraries of phenomena: a critique of Guala's methodology of experimental economics* that Guala is strongly influenced by Ian Hacking's characterization of laboratory sciences: "those whose claims to truth answer primarily to work done in the laboratory" (Hacking 1992). This influence, Jones argues,

leads Guala to overemphasize the difficulties of bridging *the gap* between internal and external validity. Guala makes a rather strict divide between testing for robustness (according to him this is an acceptable laboratory procedure) and testing for external validity (which he claims is impossible, due to the fact that experimenters cannot exactly reproduce the real system in the laboratory). This analogy between the natural and the social sciences is, however, easily drawn too far.

Many naturalistically inclined methodologists and researchers want to point out that there is an essential difference between the natural and social world with regard to the way the study objects are affected by different contexts. In fact, and this is our argument, this interplay is one of the things that threaten the priority claim of internal validity.

Social scientists worry that participants bring their experience of the world outside the laboratory with them into the experimental setting, and this may fundamentally change the way that the "target system" and the laboratory "reconstruction" of this system relate. What the researchers find to be internally valid results might be strongly dependent on them being externally valid, in a loose sense. We know them to hold outside the laboratory, and that is why we discover them in the laboratory. Furthermore, applied researchers within this field do not remain with the laboratory setting, something that further complicates the internal/external validity distinction. For instance, Baruch Fischhoff (1996), in the wonderfully titled "The real world: What good is it?", published in *Organizational Behavior and Human Decision Processes*, has argued that applied psychology can change the way that experimental psychology is conducted by allowing researchers to better understand the nature of the laboratory tasks. In particular, a little applied psychology may open researchers' eyes to "the curse of context" (that participants bring their own understanding to the minimalist problems set before them in the laboratory) and the "curse of cleverness" (devising complex experimental tasks with the assumption that participants immediately will understand their structure). The curse of context

clearly threatens the internal/external validity distinction by putting into question whether we can isolate that which we observe from the context. This might not be Fischhoff's worry, but it applies to the problem at hand. Fischhoff's mission is another. The standardization of stimuli in controlled laboratory settings turn participants into "battery raised hens", Fischhoff claims, with the hope of being able to produce predictable changes in output, whereas applied psychology studies free range poultry instead. Fischhoff's hope is that the combination of the two will lead to a better understanding of cognitive processes.

5. EBM and Vetenskap och beprövad erfarenhet (Science and proven experience)

Interestingly some of the discussions within the philosophy of natural science and economics have carried over to the more applied field of evaluating the strengths and weaknesses of evidence-based medicine (EBM) and health care. Of particular interest in this connection might be to study concepts such as (the distinctly Swedish notion) "vetenskap och beprövad erfarenhet". Nils-Eric Sahlin recently acquired a substantial amount of money from Bank of Sweden Tercentenary Foundation, and we very much hope that that programme will shed light on the distinction between internal and external validity as well. We have started to develop some such ideas in *Vetenskapsteori för sanningssökare* (Fri Tanke 2013).

Randomised controlled trials (RCT) are often seen as the privileged route to causal inference in EBM. RCTs are important in this context since they enforce both the idea that internal validity is prior to external validity and that there is a trade-off between the two types of validity. However, we should perhaps take care to distinguish causal inference from inferences involving the elimination of alternative explanations. Hence an implication to be explored emerges from a position where it is accepted that it is not a coincidence that A and B occur together and where it remains an open question if A and B are causally related. This possibility leaves open that internal validity (A causes B in the trial) depends on the external validity of the claim

(A causes B outside the laboratory). It is interesting in this context that later Campbell proposed the abandonment of the concept of internal validity and suggested 'local molar validity' (i.e. inference to a complex package of potential difference-makers) in its stead.

6. Artefacts and internal validity

The importance of how participants adapt to, and utilize, contingent features of their everyday life and bring this with them to the laboratory setting has been discussed within cognitive psychology even prior to Campbell's notion of external validity. It is often traced back to Egon Brunswik's perception research which challenged the Gestalt psychologists' focus on perceptual illusions by demonstrating a surprising degree of perceptual accuracy under natural conditions. Brunswik's insistence on performance in the natural world presupposes external validity, and has given rise to the probabilistic view on judgment and decision making that we will explore more closely below. One of his main interests was how well factors imperfectly related to a criterion to be predicted function in real life. He is, for instance, well known for research in which he tried to determine to what extent retinal size could be used to predict the actual size of an object. In principle, retinal size is not a good cue for actual size, since both objects' size and their distance to participants can vary. In practice, however, objects tend to be of certain sizes and be looked at, at certain distances. Such contingent relations can, and to some extent do, make retinal size a good cue for actual size in natural environments. Brunswik is, however, not only historically important. His emphasis on representative sampling is also a tool for identifying the situations or environments in which decision making is supposed to succeed. In his *Distal focussing of perception: Size constancy in a representative sample of situations* (Psychological Monographs, 1944), Brunswik attempted to randomly sample instances in which participants spontaneously looked at objects in their everyday life, and measure the correlation between retinal size and object size in these particular situations. The environment in which the pre-

dictive potential of retinal size is measured is thus determined through *representative sampling*.

Representative sampling is a key phenomenon in the internal/external validity debate since it emphasizes that good experimental data only can be found if the experiment is – in important respects – similar to the everyday surroundings of participants. Within, in particular, judgment and decision making research, the ideas of representative sampling have resurfaced through a relatively recent debate regarding the validity of a number of experimental findings allegedly demonstrating the inaccuracies of human judgment. The key role of external validity here is thus not to guarantee the generalizability of experimental findings (the role still exists though). Rather, the potential generalizability of the findings is what guarantees that the experimental results are not merely artefacts. This might happen both in obvious and more oblique ways.

The most obvious example is that researchers may, in the experimental task, use (or interpret) words in a way that is unfamiliar to participants, or at least different, from how participants use them. For instance, when participants are asked to state their probabilistic beliefs, 50% (.5, or similar) has an elevated frequency, presumably because phrases such as "fifty-fifty" are taken to represent uncertainty rather than a particular probability, as was established in the paper *Fifty-Fifty = 50%?* (Fischhoff & Bruine de Bruin, 1999). Sometimes differences in terminology have been argued to be the true cause of well-known experimental effects. With respect to the conjunction fallacy (related to the famous Linda-problem), it has repeatedly been argued that the fallacy is due to participants' (mis) understanding of "probability" when participants are given the task to rank statements "by their probability", or of the operator "and" when participants then rate the critical statement "Linda is a bank teller and an active feminist" to be more probable than "Linda is a bank teller".

There are, of course, many more examples, but the main point is that potential experimental artefacts such as these demonstrate that participants bring their knowledge of the surrounding world into the laboratory. In so far as the

experiment, or the experimental stimuli, in some important respect misrepresents participants' experiences, it is likely that the behaviours observed in the laboratory are mere artefacts. In these cases, both internal and external validity are compromised. Truly internally valid results require that we see clearly, i.e. that what we see in the local environment is not in fact an artefact of something else. And to be able to identify the experimental artefacts, we need to be able to see what participants see – a skill that can be trained through applied research, according to the argument of Fischhoff above.

7. Two problems

From the above one can argue that the claim that internal validity is prior to external validity is too simplistic by pointing to two epistemologically problematic aspects: experimental artefacts and the implication of causal relations. Each demonstrates how important external validity is to the internal validity of the experimental result.

For instance, 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 is internally valid. Egon Brunswik puts it neatly: "psychology has forgotten that it is a science of organism-environment relationships, and has become a science of the organism" (Brunswik, 1957, 6).

IMAGE CREDITS
Cover image: *Mein Kindermädchen*, 1936
© Meret Oppenheim/Bildupphovsrätt 2015
The British Library (p. 157), The British Museum (121), J. Paul Getty Museum (136, 138, 140), The Morgan Library & Museum (137), The National Gallery (143), Wikimedia Commons (114, 116, 118, 123, 125, 128, 134, 142, 145, 147, 149, 154)

OTTO NEURATH
Gesellschaft und Wirtschaft (p. 153) can be downloaded at medienphilosophie.net/neurath/Gesellschaft_und_Wirtschaft_1931.pdf

Fri tanke förlag
www.fritanke.se
info@fritanke.se

Design: Johan Laserna
Printed by Media-Tryck, Lund 2015
ISBN 978-91-87935-37-4