Can there be a trade-off between internal and external validity?

By Martin Vaeth

Abstract:

Proponents of randomised controlled trials (RCTs) see in them the potential to revolutionise empirical methods in the social sciences. RCTs are experiments that randomly allocate participants into two groups, one that receives a treatment (the treatment group) and one that receives no treatment or a placebo (the control group). This design extracts the causal effect of the treatment. RCTs can be compared to other empiricals methods along the two dimensions of internal validity, i.e. how good they are at finding the true causal relationship in the studied population, and external validity, i.e. how good the discovered causal relationship can be generalised to other populations. While there is broad consensus that RCTs are better in internal validity than orthodox empirical methods such as regressions, there is much controversy whether this advantage comes at a cost to external validity. Even more, there are competing views about the relationship between internal and external validity that revolve around the question whether there can be a trade-off between them. In this paper, I use a formal approach to define internal and external validity and show that a trade-off is conceptually possible and how it might arise in practice.

Recent decades have seen a rise in new experimental methods such as randomised controlled trials (RCTs) in economics and other social sciences. RCTs promise to lead to better causal inferences than orthodox methods such as regression. Duflo et al., for example, write (2004: 8) that "creating a culture in which rigorous randomised evaluations [RCTs] are promoted, encouraged, and financed has the potential to revolutionise social policy during the 21st century, just as randomised trials revolutionised medicine during the 20th." There has been criticism, however, that the power of RCTs is being overestimated due to their lack in external validity (Cartwright 2007: 12). It is often held that there is a trade-off between internal and external validity in empirical methods (ibid: 11). According to this view, RCTs may have higher internal validity than methods such as regressions but RCTs have lower external validity. A contrasting position is that internal validity is a prerequisite for external validity (Lucas 2003: 248; Guala 2003: 1198; Hogarth 2005: 262). In this case, there can be no trade-off because it is not possible that external validity increases while internal decreases. I think these contrasting views result from a lack of clarity about what internal and external validity are and what their relationship is.

To clarify the matter, I propose a formal definition of internal and external validity. Using these definitions, I show that a trade-off between internal and external validity is possible and how it might arise. Section 1 introduces my notion of causality and gives formal definitions of internal, external and overall validity of an empirical method. These definitions will be measures of how close a coefficient is to the real causal coefficient, namely a mean squared error. Section 2 applies the formal definitions to a hypothetical example comparing an RCT with a regression and shows how the regression might be better in external but worse in internal validity than the RCT. I essentially give a formalisation of the idea that a regression can be better in external validity because it is based on a more heterogeneous sample.

Section 1

For the purpose of this article, I understand causality as a probabilistic relation between variables. Variable A having a causal effect on variable B means that if all other variables stay the same, a change in A leads to a change in the probabilistic distribution of B (see Hitchcock 1997: section 2). For example, A could be the price of a distributed insecticide-treated bed net in sub-Saharan Africa and B the usage of that bed net or the effect on malaria prevention, to give an example from the literature (Cohen and Pascaline, 2010). The causal effect of A on B can depend on the value of other variables and therefore be different for different regions and it can also depend on the value of A. To simplify the following analysis, I suppose the causal effect of A on B, given the

values of the set of other variables, V, has a linear form:

$B = f(V) + \beta(V) \cdot A$

This means that B is function of V plus a coefficient β that may depend on the variables V times the value of A. I use this simplification because it allows us to capture the causal effect in a single coefficient β and because such a form is commonly assumed in practice.

For the rest of this article, I suppose that the purpose of an empirical method is to find the true causal coefficient of one variable on another variable in a 'target population'. An empirical method does so by studying a population which I will call 'population of study'. According to the classical definitions by Cook and Campbell (1979: 37), internal validity "refers to the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause", and external validity "refers to the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings, and times." Let us call β the causal coefficient that our empirical method yields, β_1 the true causal coefficient of the population of study and β_0 the true coefficient of the target population. Intuitively, internal validity should be a measure of how close β is to β_1 and external validity should be a measure of how close β_1 is to β_0. When measuring the proximity of these coefficients, we should keep in mind that β is a random variable (and later I will argue that β_1 and β_0 can be seen as random variables as well). A first measure of internal validity would be the bias $\mathbf{E}(\beta - \beta_1)$. However, even if the bias is very small, β can usually be very far from β_1 if the variance of β is large. A better measure for internal validity that takes into account both the variance of our estimator and the bias is the mean squared error $\mathbf{E}((\beta - \beta_1)^2)$ which is commonly used in probability theory.

Internal Validity: $\mathbf{E}((\beta - \beta_1)^2)$

It can be shown that the mean squared error is exactly the sum of the squared bias and the variance of β. The mean squared error is always non-negative. If it is zero, we find exactly the right coefficient. The bigger it is, the less precise is our estimator β, i.e. the weaker is our internal validity.

Like written above, external validity is generally understood as a measure of how good the causal effect can be generalised from the population of study to the target population. That is why I take external validity as how close the coefficient of the population of study, β_1, is to the coefficient of the target population, $\beta_{0.}$ This does not depend on the empirical method itself (it is independent of internal validity) but on how similar the two populations are. So I define external validity analogously to above as:

External Validity: $\mathbf{E}((\beta_1 - \beta_0)^2)$

Finally, I define overall validity as the as the expectation of the squared distance of our estimator β and the true coefficient of our target population β_0:

Overall Validity: $\mathbf{E}((\beta - \beta_0)^2)$

These definitions acknowledge the fact that we can never find the true coefficient with perfect precision but that we only have estimates. While we are using probabilistic expectations, we do not establish whether the underlying probabilities are objective or epistemic. My formal definition of validity allows both interpretations.

Now we can turn to the connection between internal, external validity and overall validity. We can mathematically derive the following equation (called equation (1) from now on):

$$\mathbf{E}((\beta- \beta_0)^2) = \mathbf{E}([(\beta- \beta_1) + (\beta_1 - \beta_0)]^2) = \mathbf{E}((\beta- \beta_1)^2) + \mathbf{E}((\beta_1 - \beta_0)^2) + 2 \cdot \mathbf{E}((\beta- \beta_1)(\beta_1 - \beta_0)) \qquad (1)$$

Here I used that $\mathbf{E}(A+B)= \mathbf{E}(A)+ \mathbf{E}(B)$ for random variables A and B. (1) shows that overall validity is the sum of internal validity, external validity and a third term. This third term can either be positive (weakening overall validity) or negative (strengthening overall validity). It strengthens overall validity if it happens that β overestimates β_1 but (incidentally) β_1 is smaller than β_0, thus β ends up being close to β_0. The third term captures this combination effect that even poor external and internal validity can theoretically result in strong overall validity. As we cannot depend on this combination effect to improve overall validity, we can see from (1) that we need both a strong internal and a strong external validity for a strong overall validity. Furthermore, (1) shows that the role that internal and external validity play in determining overall validity is symmetric. Thus, a trade-off is conceptually possible.

Section 2

To see how there could be a trade-off between internal and external validity in practice, let us look a hypothetical example. In this example, a regression will have a weaker internal but stronger external validity than an RCT. Let us suppose that we want to find the causal effect of the price of insecticide-treated bed-nets on their usage. These bed-nets are distributed by aid programmes to prevent malaria. According to one hypothesis, people value the bed-nets more if they paid something for them and therefore use them more. Our target population is region X in sub-Saharan Africa where we want to know the causal coefficient. We have an RCT in another region Y in sub-Saharan Africa and a statistical regression from data of regions $Z_1, \ldots, Z_n$ in sub-Saharan Africa. They yield the coefficients β_{rct} and β_{reg}, respectively. The true coefficient of region Y is $\beta(Y)$, the average coefficient of regions $Z_1, \ldots, Z_n$ is $\beta(Z)$ and the true coefficient of the target population region X is β_0.

It is commonly held that RCTs tend to have a stronger internal validity than regression analyses. The basic reason is that in a RCT the experimenter can select the control group and treatment group randomly while in a regression they cannot be chosen. In the latter case, different forms of selection bias might arise. Thus, I suppose that the internal validity of the RCT is stronger than the one of the regression.

My main aim is to argue that in some cases we have reason to believe that the external validity of a regression is stronger than the one of an RCT. This can be the case because regressions typically base their estimate of the causal coefficient on a more heterogeneous population than RCTs. Let us suppose we could divide sub-Saharan Africa into regions of equal size such that each region has one homogeneous causal structure (and thus one causal coefficient).8 This is of course wrong, but that does not matter as this example is purely hypothetical. Furthermore, we assume that we have no information at all about whether $\beta(Y)$, $\beta(Z_i)$ and β_0 are at the high or low spectrum of the distribution of causal coefficients in different regions. To say something about the external validity I make a crucial step: We can think about the coefficients β_0, $\beta(Y)$ and $\beta(Z_i)$ (for every region Z_i in Z) as drawn randomly and independently from the distribution of coefficients in regions of Africa. Thus, they all have the same distribution and especially, same expected value and same variance σ^2. We should interpret this not as an assumption about objective probabilities of the values of these coefficients but as one about epistemic probabilities. In the context of external validity, objective probabilities are hard to apply as our choice of the target population and the population of study might not be the result of a random experiment. We probably choose them according to some criteria like availability of data or suitability for a RCT. As long as we do not believe these criteria are correlated with the causal coefficients in the region, we can motivate the assumption above by a principle of indifference or a flat prior Bayesian belief. Then we have the following external validity of the RCT:

$$\mathbf{E}((\beta(Y) - \beta_0)^2) = \mathbf{E}(\{[\beta(Y) - \mathbf{E}(\beta(Y))] + [\mathbf{E}(\beta_0) - \beta_0]\}^2) =$$

$$\mathbf{E}([\beta(Y) - \mathbf{E}(\beta(Y))]^2) + \mathbf{E}([\mathbf{E}(\beta_0) - \beta_0]^2) + 2 \cdot \mathbf{E}([\mathbf{E}(\beta(Y)) - \beta(Y)][\mathbf{E}(\beta_0) - \beta_0]) =$$

$$\mathrm{Var}(\beta(Y)) + \mathrm{Var}(\beta_0) + 2 \cdot \mathbf{E}((\mathbf{E}(\beta_0) - \beta_0)) \cdot \mathbf{E}(\mathbf{E}(\beta(Y)) - \beta(Y)) =$$

$$\mathrm{Var}(\beta(Y)) + \mathrm{Var}(\beta_0) = 2 \cdot \sigma^2$$

The first equality holds because $\mathbf{E}(\beta(Y)) = \mathbf{E}(\beta_0)$ and the third because β_0 and $\beta(Y)$ are independent.

Analogously, we obtain the following external validity for the regression:

$$\mathbf{E}((\beta(Z) - \beta_0)^2) = \mathrm{Var}(\beta(Z)) + \mathrm{Var}(\beta_0) = \mathrm{Var}(\beta(Z)) + \sigma$$

[8] This means that we assume there is no unmodeled causal heterogeneity. If such unmodelled causal heterogeneity was present, then additional problems with the internal validity of the regression would occur (see Aronow and Samii, 2016).

To determine Var(β(Z)), we first remember that because the regions are of equal size, the true causal coefficient of the set of regions Z is the average over the coefficients of the different regions in Z, so $\beta(Z) = \frac{1}{n}\sum_{i=1}^{n} \beta(Z_i)$. Because β(Z) is an average, it has a lower variance than each $\beta(Z_i)$:

$$\mathrm{Var}(\beta(Z)) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} \beta(Z_i)) = \frac{1}{n^2}\mathrm{Var}\ (\sum_{i=1}^{n} \beta(Z_i)) = \frac{1}{n}\sigma^2$$

Here I used that $\mathrm{Var}(cA) = c^2\,\mathrm{Var}(A)$ and $\mathrm{Var}(A+B) = \mathrm{Var}(A)+\mathrm{Var}(B)$ for any constant c and independent random variables A, B. This means that the external validity of the regression is smaller than the one of the RCT:

$$\mathbf{E}((\beta(Z) - \beta_0)^2) = (1 + \frac{1}{n})\,\sigma^2 < 2\cdot \sigma^2 = \mathbf{E}((\beta(Y) - \beta_0)^2)$$

This effect can be interpreted as follows: The RCT looks only at one region while the regression averages over many regions. It is less probable for the average over many regions than just for the coefficient of one single region to be far from the average coefficient in sub-Saharan Africa.

Finally we have to consider the effect of our empirical method on the third term in overall validity, which is $2\cdot\mathbf{E}((\beta-\beta_1)(\beta_1-\beta_0))$. Here β stands for the coefficient that the RCT or the regression yields, respectively, and β_1 stands for β(Y) or β(Z), respectively. It is reasonable to assume that the internal error $\beta-\beta_1$ and the difference $\beta_1-\beta_0$ are independent in both cases, so $\mathbf{E}((\beta-\beta_1)(\beta_1-\beta_0)) = \mathbf{E}(\beta-\beta_1)\cdot\mathbf{E}(\beta_1-\beta_0)$. Because $\mathbf{E}(\beta(Y)) = \mathbf{E}(\beta(Z)) = \mathbf{E}(\beta_0)$, the second factor and hence the whole term is zero in both cases.

To sum up, I have proposed a formal definition of internal, external and overall validity. Not only do these definitions capture the intuition behind these terms in an adequate way, they prove to be fruitful concepts to analyse the validity of an empirical method. In the example in section 2, I made a number of assumptions to isolate the effect that a regression can have a stronger external validity than an RCT because it is based on a more heterogeneous sample. The strength of the formal analysis is to highlight all the assumptions that need to hold for this effect and to hint at other possible effects that might affect validity.

References

Aronow, P. M., & Samii, C. (2016). Does regression produce representative estimates of causal effects?. *American Journal of Political Science*, *60* (1), 250-267.

Cartwright, N. (2007). Are RCTs the gold standard?. *BioSocieties, 2* (1), 11-20.

Cohen, J., & Dupas, P. (2010). Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics,* 1-45.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Boston: Houghton Mifflin.

Duflo, E., Glennerster, R., & Kremer, M. (2004). Randomized evaluations of interventions in social service delivery. *Development Outreach, 6* (1), 26-29.

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

Hitchcock, C. (1997). Probabilistic causation. *The Stanford Encyclopedia of Philosophy.* Winter 2016 Edition. https://plato.stanford.edu/entries/causation-probabilistic/#ProRaiTheCau

Hogarth, R. M. (2005). The challenge of representative design in psychology and economics. *Journal of Economic Methodology, 12* (2)*,* 253-263.

Lucas, J. W. (2003). Theory-testing, generalization, and the problem of external validity. *Sociological Theory, 21* (3), 236-253.