Key principles of evaluation science

The ultimate goal of social science research is to produce reliable knowledge or to provide the evidence that guides applied decision-making. In pursuit of this goal, it is essential that components of the research process – including design, methods, analysis, and conclusions – are dependable. It is crucial, for instance, that methods and findings of research provide an accurate reflection of the truth. While this concept is broadly defined as research validity, there are different types and each should be considered in the process of furthering knowledge and translating findings into evidence-based policy. However, given that one of the main bottlenecks to evidence-based policy making is implementation of sound results, external validity should be emphasized as the most important lever in translating research findings in effective policy.

The concept of validity is an essential part of social science research because it indicates how well a test or method measures what it is supposed to measure. In essence, validity is an indication of how sound research is. According to Carmines et al., for example, “ an indicator of some abstract concept is valid to the extent that it measures what it purports to measure” (Carmines et al., 2008, pg. 12). Unlike reliability, which is the extent to which an experiment, test, or measurement yields the same results on repeated trials, validity is an indication of whether or not a study measures what it is supposed to measure (i. e. how accurate a study is).

Importantly, validity is a multidimensional concept that is relevant to all stages of the research process, from data collection and design to methods and conclusions. To illustrate, a recent study by Suchon et al. found that social mobility reduces trust using a controlled laboratory experiment with

university students in France (Suchon et al., 2018). While there are numerous frameworks that can be used to analyze the validity of this study, one example includes questioning whether the method the study uses to measure trust is actually a good representation of trust as a concept. Another validity-related question is asking whether the results of the study would also hold in a non-French university context. It is possible, for instance, that particularities of a French university context result in social mobility actually increasing trust in another context. As with the first question, we can gauge validity with close-inspection of the methods, design, and reasoning used by the researchers.

One framework used in analyzing research validity, called statistical conclusion validity, involves determining whether conclusions of a research study are founded on an adequate analysis of the data. The concept of statistical conclusion validity can be defined as the extent to which data from a research study can reasonably be regarded as revealing a link between independent and dependent variables. In particular, statistical conclusion validity " is concerned with sources of random error and with the appropriate use of statistics and statistical tests" (Cook and Campbell, 1979, pg. 39–50).

In essence, analyzing statistical conclusion validity of a study involves asking the question: did the investigators arrive at the correct conclusion regarding whether or not a relationship between the variables exists or the extent of the relationship? To illustrate, we can once again consider the study by Suchon et al. 2018 that investigates the relationship between social mobility and trust. One could gauge the statistical conclusion validity of this study by analyzing the sampling procedures, statistical tests, and measurement

procedures used. For example, one might question whether the data comes from a reliable source or was measured in a reliable way. Alternatively, one could consider whether the study used a large enough sample. If a small sample was used, then the study could have low statistical power and, thus, be prone to find no statistically significant effect when there actually is one (i. e. a type II error). Statistical conclusion validity is not concerned with the causal relationship between variables, for example between social mobility and trust, but whether or not there is any relationship either causal or not. Statistical conclusion validity can be strengthened in research with appropriate sampling procedures, statistical tests, and measurement procedures.

While statistical conclusion validity is a measure of how reasonable a research or experimental conclusion is, construct validity, another type of validity, is the degree to which a test measures what it claims it is measuring. In assessing construct validity, we consider the overlap between the scientific operationalization of a higher-order term (such as trust, social mobility, intelligence, etc.) and the meaning of this higher order term in practice. Consider the broad concept of intelligence which is arguably difficult to quantify and describe. An IQ test is an attempt to quantify intelligence and many studies have used IQ scores as a proxy for studying the relationship between intelligence and some variable. For example, recent work by Zajenkowski et al. investigated the relationship between per-capita GDP and the average intelligence of the population (Zajenkowski et al., 2013). If one were to assess the construct validity of that study, one might evaluate, based on research, the overlap between the scientific

operationalization of intelligence (via IQ tests) and the true meaning of intelligence as a concept.

Construct validity is particularly important when the concept used does not have a standard method of measurement. Relative to some other concepts, intelligence has a fairly standard measure (the IQ test) and has been explored in a fair amount of depth. However, there are numerous other concepts that researchers consider that need even more careful scrutiny around construct validity. For example, what is an accurate measure of principal effectiveness? Or what sort of test can measure emotional stability of young children? Construct validity is the extent to which any test used is actually measuring the construct it claims it is measuring. Concerns of construct validity are incredibly important to ensure that conclusions drawn are accurate to the reality. If the measurement of a concept in research is not well-aligned with the actual concept, then any research findings related to the concept will be inaccurate. Construct validity can be strengthened with well-established definitions and measurement procedures for variables.

Considering that a substantial portion of social science research is concerned with cause-and-effect relationships between variables, a particularly important type of validity is internal validity, which evaluates the extent to which a study can rule out alternative explanations for its findings. Internal validity is a measure of accuracy of the experiment and addresses whether or not observed covariation should be considered a causal relationship. It is a particularly important consideration in evidence-based social policy research in which, often, researchers estimate the impact of particular programs on outcomes. In these contexts, one would like to make accurate conclusions

about the causal effect of a particular treatment or program in a way that it is isolated from impact of confounding variables on the outcome of concern.

Questions related to internal validity and causal effects have been a topic of substantial work in statistics and econometrics. Donald Campbell and Julian Stanley, in their work *Experimental and Quasi Experimental Designs,* raised issues about threats to internal validity (whether or not observed covariation should be interpreted as a causal relationship) that exist when researchers are not able to randomly assign participants to treatments (Campbell and Stanley, 1966). More recently work by Donald Rubin and Guido Imbens has furthered research into estimating causal effects, with a particular emphasis on cases in which data is observational. For example, consider research into the relationship between smoking and cancer. A study that investigates the impact of smoking on cancer would have high internal validity if there are strong cause-and-effect conclusions.

The major threat to internal validity, in this research example and others, is a variable that often co-varies with the independent variable (smoking) and may be an alternative cause of the dependent variable (cancer), which is also called a confounding variable. To investigate whether smoking cigarettes causes cancer, one needs to control for other possible causes of cancer that tend to vary with smoking. For example, Fukumoto et al. identified age, sex, drinking status, fruit and vegetable intake, family history of lung cancer, occupation, and years of education as confounding variables in the relationship between smoking and lung cancer (Fukumoto et al., 2015). Internal validity is important to address or eliminate alternative explanation for a particular result, and can be improved by ensuring that

extraneous variables have been controlled and confounds have been eliminated. Typically, this can be done through good study design and rigorous analysis methods.

While internal validity focuses on the relationship between variables, external validity is concerned with the universality of the results. External validity is the degree to which the conclusions in a study would hold for other persons in other places and at other times. In evidence-based policy, external validity is a particularly important because it is often enacted based on previous research that might have occured in the same context at an earlier time or in a different context. To have a high degree of external validity is to have confidence that results of particular research will generalize to another context. A research project by Edward Miguel and Michael Kremer, for instance, found that the externalities of mass school-based deworming in Kenya were large enough to justify a full subsidy of the program (Miguel and Kremer, 2004). In response to this research, the Government of India launched a national deworming program in 2015 that now reaches 260 million school children each year (“ Deworming to increase school attendance”). Confidence in the external validity of results were essential for the findings in Kenya to be applied in India. External validity is important because it broadens the relevance of findings to contexts or populations that weren’t studied. To strengthen external validity in a study, researchers can measure dependent variables under natural conditions and broaden the context of study.

The external validity of research is dependent on its internal validity because, without valid within-study conclusions, inferences about the

general population will also be inaccurate. Internal validity can also be described as the extent to which an experiment is free from errors and whether or not observed covariation should be considered a causal relationship, whereas external validity can be described as the extent to which the research results can be inferred to the world at large. The difference between internal and external validity can most obviously be understood through the questions one might ask in assessing each. In assessing internal validity, one might ask: “ How strong are the research methods?”. In assessing external validity, one might ask: “ Can the outcome of the research be applied to the real world?”

Without internal validity, external validity becomes much less important. After all, external validity checks whether the causal relationship discovered in an experiment can be generalized or not. That causal relationship must first be established via internal validity. In other words, the conclusions of a study must be assessed before the degree to which the study is warranted to generalize the result to other contexts can be assessed. Once again, consider the deworming research by Miguel and Kremer (2004). If their methods did not result in sufficient internal validity, then they would not have been able to make strong claims about the causal effect of the school-based deworming program on outcomes such as student attendance. As a result, questions of external validity, for example regarding the generalizability of results in a Kenyan context to an Indian context, become much less important.

However, an overemphasis on internal validity relative to external validity can greatly constrain the ability of research to be applied in practice. It has

been frequently argued that internal validity is the priority for research. For example, Francisco Guala, a philosopher, claimed 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). Donald Campbell and Julian Stanley argued for a greater emphasis on internal validity in *Experimental and Quasi Experimental Designs* and, historically, there has been a tendency towards researchers maximizing internal validity (Steckler et al., 2008). In part, this emphasis is justified: there is a sense that it is more important to know whether a particular social policy or program works under highly controlled conditions than it is to know if it will work among different population groups, organizations, or settings. However, optimizing for internal validity can constrain external validity and, as a result, reduce the likelihood that findings will actually be applied via policy. According to Steckler et al. (2008), “ funding organizations and journals have tended to be more concerned with the scientific rigor of intervention studies than with the generalizability of results…[which has] contributed to failure to translate research into… practice”. Particularly in an applied discipline such as evidence-based social policy, it is important to know not only that a program is effective, but that it is likely to be effective in other settings and with other populations.

External validity is more important than internal validity because it presents a greater bottleneck to evidence-based policy making. There are a growing number of impact evaluations worldwide, many of which identify effective policies and programs. According to Dr. Martin Williams, “ the question of

how to apply this evidence in policy making processes has arguably become the main challenge for evidence-based policymaking" (Williams, 2017). And there is a well-documented lag, which Bellamy et al. called ' egregious', between research development and the application of findings to practice. A greater emphasis on external validity – via a broad context of study or, at the very least, a thorough, explicit evaluation of external validity in research materials – has the potential to greatly reduce that lag. While internal validity is obviously a framework that should not be completely neglected, given the current state of evidence-based social policy, its emphasis should be reduce and instead placed on external validity. Essentially, external validity is a lever with a greater potential to increase evidence-based policy making by making it easier for policy-makers to confidently adopt findings from one context and apply them to their own context.

A common criticism of academic research is that, even if compelling research is produced, there is a relatively low likelihood that the findings will be applied in practice or that they will be applied in a timely manner. Concerns of validity are central to the research process and provide frameworks for thinking about whether the conclusions put forth are an accurate reflection of reality. Four of the most common validity frameworks include statistical conclusion validity, construct validity, internal validity, and external validity, and each help ensure that research is dependable. However, in an applied discipline such as evidence-based social policy, it is particularly important that external validity be emphasized and strengthened, so that findings from one context can be confidently applied in

a different context. In doing so, the barriers to translating research into social policy practice will be greatly reduced.

References

- Bellamy, J. L., Bledsoe, S. E., & Traube, D. E. (2006). The Current State of Evidence-Based Practice in Social Work. *Journal of Evidence-Based Social Work, 3* (1), 23-48. doi: 10. 1300/j394v03n01_02
- Bryman, A. (2001) Social research methods. Oxford: Oxford University Press
- Campbell DT, Stanley JC. *Experimental and Quasi Experimental Designs* . Chicago, Ill: Rand McNally; 1966.
- Cook T. D., Campbell D. T. (1979). Quasi-Experimentation: Design and Analysis Issues for Field Settings. Boston, MA: Houghton Mifflin
- Carmines, E. G., & Zeller, R. A. (2008). Reliability and validity assessment. Newbury Park, CA: Sage Publ.
- Deworming to increase school attendance. (n. d.). Retrieved from https://www. povertyactionlab. org/case-study/deworming-schools-improves-attendance-and-benefits-communities-over-long-term
- Field, A. & Hole., G. (2003) How to design and report experiments. London: SAGE
- Fukumoto, K., Ito, H., Matsuo, K., Tanaka, H., Yokoi, K., Tajima, K., & Takezaki, T. (2015). Cigarette smoke inhalation and risk of lung cancer. *European Journal of Cancer Prevention, 24* (3), 195-200. doi: 10. 1097/cej. 0000000000000034
- Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science* , 70(5): 1195-1205

- Imbens, G. W., & Rubin, D. B. (2016). *Causal inference for statistics, social, and biomedical sciences: An introduction* . Cambridge: Cambridge University Press.
- Miguel, E., & Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica, 72* (1), 159-217. doi: 10. 1111/j. 1468-0262. 2004. 00481. x
- Roe, B. E., & Just, D. R. (2009). Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments, and Field Data. *American Journal of Agricultural Economics, 91* (5), 1266-1271. doi: 10. 1111/j. 1467-8276. 2009. 01295. x
- Shadish, W., Cook, T. D., Campbell, D., (2002) Experimental and quasi-experimental designs for generalised causal inference. Boston : Houghton Mifflin
- Steckler, A., & Mcleroy, K. R. (2008). The Importance of External Validity. *American Journal of Public Health, 98* (1), 9-10. doi: 10. 2105/ajph. 2007. 126847
- Suchon, R., & Villeval, M. C. (2018). Does Upward Mobility Harm Trust? *SSRN Electronic Journal.* doi: 10. 2139/ssrn. 3100711
- Williams, M. J. (2017). External validity and policy adaptation: From Impact Evaluation to Policy Design (BSG Working Paper Series No. BSG-WP-2017/019).
- Zajenkowski, M., Stolarski, M., & Meisenberg, G. (2013). Openness, economic freedom and democracy moderate the relationship between

national intelligence and GDP. *Personality and Individual Differences, 55* (4), 391-398. doi: 10. 1016/j. paid. 2013. 03. 013