GENERALIZING CAUSAL KNOWLEDGE IN THE POLICY SCIENCES: EXTERNAL VALIDITY AS A TASK OF BOTH MULTI-ATTRIBUTE REPRESENTATION AND MULTI-ATTRIBUTE EXTRAPOLATION

Thomas D. Cook

INTRODUCTION AND PURPOSE

I have been asked to write about methodological issues likely to be prominent in future public policy research. Many issues would deserve attention in a longer presentation, but here I want to concentrate on external validity and its links to evidence-based policy. Such policy uses social science knowledge about what has worked in the past to inform policy decisions in the future. This requires justified procedures for describing the populations of persons, settings, and times in which a given causal relationship has been demonstrated to date, and justified procedures for moving from operational details about the cause and effect to the category labels we use to designate the more general cause or effect constructs. We call this the *representation function* since the need is to know what the sampling particulars represent as more general populations or categories. The traditional sampling theory framing of this issue would be the following: Given the populations of persons, settings, times, and treatment and outcome constructs to which I want to generalize, how well do the specifics actually sampled match these populations or categories?

Informing future policy decisions also requires justified procedures for extrapolating past findings to future periods when the populations of treatment providers and recipients might be different, when adaptations of a previously studied treatment might be required, when a novel outcome is targeted, when the application might be to situations different from earlier, and when other factors affecting the outcome are novel too. We call this the *extrapolation function* since inferences are required about populations and categories that are now in some ways different from the sampled study particulars. Sampling theory cannot even pretend to deal with the framing of causal generalization as extrapolation since the emphasis is on taking observed causal findings and projecting them beyond the observed sampling specifics.

We argue here that both representation and extrapolation are part of a broad and useful understanding of external validity; that each has been quite neglected in the past relative to internal validity—namely, whether the link between manipulated treatments and observed effects is plausibly causal; that few practical methods exist for validly representing the populations and other constructs sampled in the existing literature; and that even fewer such methods exist for extrapolation. Yet, causal extrapolation is more important for the policy sciences, I argue, than is causal representation.

THE FIVE ATTRIBUTES OVER WHICH CAUSAL REPRESENTATION AND EXTRAPOLATION ARE NEEDED

Cronbach (1982) has described one framework for understanding the generalization of causal relationships. It is schematized below, with a few changes. In this terminology, lower case u, t, o, s, and ti refer to *sampling* specifics. The u designates the

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

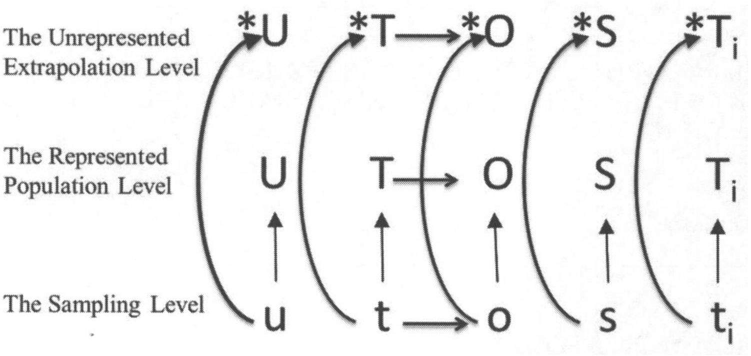
The Unrepresented Extrapolation Level
The Represented Population Level
The Sampling Level
*U
*T
*O
*S
*T_i
U
T
O
S
T_i
u
t
o
s
t_i

Figure 1. A Model of External Validity as Both Representation and Extrapolation.

various *u*nits of people or aggregates used in a study as treatment providers or as respondents; t refers to the *t*reatment actually manipulated whose effects are under examination; o refers to the *o*utcomes or effects as they are actually measured; s designates the *s*ettings in which data collection actually takes place; and ti refers to the *ti*mes (or historical moments) in which a study or set of studies has taken place. Internal validity is about whether the t to o link is causal. Figure 1 illustrates this by the horizontal arrow at the level of study sampling specifics. Much discussion has recently been devoted to methods for inferring cause, with particular emphasis coming to be placed on random assignment, regression-discontinuity, and sophisticated matching, and their potential for meeting the theoretically crucial strong ignorability assumption—also called the hidden bias or conditional independence assumption.

Before considering the rest of the figure, it is worth noting that Cronbach's framework implies two crucial insights about the t to o causal link that are often overlooked. First, if causal statements are to be useful, they require labeling the many operational details involved within the t manipulation or measure and within the o measure. Short-hand cause and effect labels that summarize these details are needed to create the T and O category labels that researchers use in their speech. What should these general cause and effect labels be is the issue, for without such short-hand it is impossible to talk sensibly about internal validity? How can one discuss whether the relationship between t and o is causal if one cannot use t and o to validly name T and O? Actually, the situation is even more difficult than this, since the causal agent is not the target T designated by a category label, but is instead the contrast between what takes place in the treatment group and whatever comparison groups are used.

Second, and perhaps even more important, is that t and o cannot occur without the sampling particulars in u, s, and ti. Any of these might condition the t to o relationship and render incomplete any conclusion about t as the sole cause, because no consideration is given to the roles that u, s, and ti might have played in bringing about the effect. Mackie's (1974) theory of INUS conditionals considers the treatment to be an *i*nsufficient but *n*onredundant part of an *u*nnecessary but *s*ufficient cause of an effect. This view assigns to t the role of an initiating causal agent within a broader and more multiattribute explanatory constellation of factors that also includes u, s, and ti. Even in a randomized experiment where u, s, and ti are balanced between groups, the intent-to-treat estimate still conflates direct effects of the intended treatment with any interactions between t and attributes of the persons studied, of the situations sampled, and of the historical period examined. Wherever such causal heterogeneity is suspected, replicated results are less likely, even in experiments with similar treatment manipulations (t) and outcome measures (o).

EXTERNAL VALIDITY TASK: I. REPRESENTATION

In traditional statistics, the aspiration is to formulate clearly the populations of units, treatments, outcomes, settings, and times to which generalization is sought, and then to sample utosti specifics that aim to represent these clearly labeled *populations*, *universes, categories, or target*s of inference, call them what one will. The ascending arrows from utosti to UTOSTi in Figure 1 illustrate how sampling specifics are meant to represent prespecified populations or categories in each of the five domains. For the links from u to U and from s to S, random sampling provides the best external warrant for claims about what the sampling specifics represent at the population level. However, random sampling is very rare in experiments and in most quasi-experiments; opportunistic samples are the rule, and category membership is usually the criterion by which population claims are validated. Yet, while a sample of, say, four-year-olds comes from within the category of that name, in most research the four-year-olds will have been opportunistically selected at some sites and so will not be formally representative of all four-year-olds in a nation, say. Moreover, the sampled children (u) are likely be further biased because their site managers and parents volunteered them to be in the study. Yet, the target U is not likely to include volunteering as a relevant category attribute. A propensity score-matching method has been proposed for generalizing from u to U, but the method depends on large samples, on very well-defined population and sampling particulars, and especially on considerable overlap on observables between the sample (u) and the intended target (U) (Stuart et al., 2011; Tipton, 2013). So, this propensity score method is a satisficing method for promoting representativeness. It is not optimal because it does not involve formal sampling, to date it applies in only one of five generalization domains, and it is largely irrelevant to extrapolation tasks within causal generalization.

For the ascending arrows from t to T and from o to O, best practices from psychometrics emphasize the following. First, the theoretical explication of a cause or effect category; then, discriminating each category from the known cognates with overlapping attributes; next, selecting instances to represent the target and the cognate categories; and finally doing construct validity tests that have convergent and discriminant validity components. Within the framework of multitrait, multimethod matrices, such construct validation entails demonstrating convergent validity across different manipulations or measures of a target construct and then discriminant validity between contending cognate concepts. Such careful practice in the validation of t and o measures is rare in causal research. Instead, practitioners resort to a satisficing alternative predicated on *surface similarity*. To illustrate, if I have an explicit theory of teacher training and my manipulation of it looks like that theory, then I assert that I have manipulated teacher training. Yet, my implemented version of teacher training is bound to include many components that (1) are not part of teacher training per se (e.g., where it takes place), (2) have fewer specific values in applications when compared to theory (e.g., how long the treatment lasts and who does the training of teachers), and (3) cannot include all of the attributes of all reasonable theories of teacher training so that *teacher training* as operationalized is but a specific version of the more general category. Ditto for the relationship between theoretical and operational definitions of, say, *academic achievement*. The currently dominant practice of *surface similarity* is intuitive, but it falls below recommended state-of-the-art practices in construct validity, such as multitrait multi-method matrices.

The time domain is especially debatable for causal generalization. How can any one time or set of times in the past (ti) be extrapolated to future times (*Ti)? We can commit the contextually minor sin of induction and assume that a stable causal relationship from the past will continue in the future, especially if the causal relationships have been multiply replicated across heterogeneous instances within each UTOS domain. Unfortunately, many causal relationships are not very stable across past studies purporting to represent even the same UTOS, and the obtained replication pattern may not follow a clear and systematic temporal sequence. Either of these circumstances makes it more difficult to draw valid conclusions about the time periods over which a causal relationship can be represented and about what to expect in the future.

When social scientists talk about the populations represented in causal studies, they seem to feel comfortable making such generalizations without using the best representational methods currently available—namely, random selection and full construct validation. Instead, practice has drifted toward accepting purportedly *satisficing* suboptimal procedures for the random selection and full construct validation procedures that are missing. These satisficing procedures include the propensity score techniques mentioned earlier, surface similarity that equates representativeness with category membership, and replication across heterogeneous populations and categories within or across studies. This taste for satisficing methods on the representational side of external validity is not matched when it comes to internal validity, inferring whether the t to o link is causal. Here the push is toward accepting only the currently optimal methods, especially random assignment even though it often entails suboptimal tests of how t and o relate to T and O and all too rarely explores how T to O relationships depend on different factors within U, S, and Ti. Ways exist to improve the representativeness of causal knowledge from experiments and quasi-experiments, but they are not currently assigned a high value, given the satisficing procedures currently in vogue.

EXTERNAL VALIDITY TASK: II. EXTRAPOLATION

As difficult as representation is when causal hypotheses are at play, extrapolation is even more difficult. Yet, as we will see, it is even more important than representativeness as a goal for public policy research that seeks to identify manipulable causal agents that can be introduced into practice at scale. In Cronbach's schema, extrapolation is represented by *UTOSTi (spoken as star-UTOSTi) and the vertical arrows to it from the utosti sampling particulars point to the need to draw conclusions about persons, settings, times, and treatment and outcome variants that have manifestly different attributes from those observed at the sampling level. At issue is drawing conclusions about how past causal relationships would hold up with populations or categories different from those previously observed. And to make matters worse, any one of the five generalization targets in *UTOSTi might share many, some, or even no attributes with the previously observed utosti details.

The epistemological problem can be simply illustrated with reference to *U. Imagine an existing archive describing a causal finding from studies with four-year-olds. It is very unlikely that the archive will include studies from all the locations where potential users of the available information work who might want to apply the causal knowledge to the particular four-year-olds for whom they are responsible; and it may not include reliable information about four-year-olds who are gifted, or whose parents are rich, or who do seasonal agricultural labor. And the persons whose preschool teaching is included in the archive might not include recent university graduates volunteering to teach for a year, or teachers who refuse permission for their class to be tested. The same is true for *T. Variants of T might have been observed in the past, but they collectively fail to include local adaptations of T that often take place once a treatment is no longer in its developer's hands. Or a potential user of the information might be interested in a version of prekindergarten that includes much more playtime than recent curriculum-centric approaches allow, or that emphasizes self-regulation more than cognitive learning. We could repeat this with *O, *S, and *Ti. The point is that future policy is almost always about options not fully explored in the existing literature. And even if satisficing information exists about what the utosti sampling details represent at the UTOSTi level, many potential users of the information–and almost all local users–would still want to know whether a given intervention will work with their clients, in their settings, at their desired time of application, with their adaptations of treatment, and with outcomes different from what is in the existing literature. So, extrapolation is intrinsic to all evidence-based policy research.

The aspiration is always to use causal evidence from the past to predict what will work in the immediate future. But at this later time, novel features might affect the size or direction of the past causal relationship in question. This predicament is universal across all the sciences, but may be exacerbated in social policy worlds when compared to natural science ones due to the less secure and more conditional nature of the knowledge being applied, to the more open systems in which the knowledge will be applied, and the sometime need to adapt past interventions rather than adopt them faithfully in accordance with some theory. If Cronbach is correct that a key task of the applied social sciences is to extrapolate causal knowledge to novel *UTOSTi policy circumstances, then the questions arise: How do we do this now, and how might we do it better in the future?

CURRENT METHODS FOR EXTRAPOLATING CAUSAL KNOWLEDGE

In many natural science contexts where it is normative (and easy) to replicate causal findings across independent laboratories, standard practice seems to be *to obtain one independent replication of a causal finding and then to assume the relationship is stable until proven otherwise*. Disconfirming proof would come from a failure to replicate under conditions very much like the original ones, at which point testable hypotheses can emerge about the contingencies affecting relationship size. Little about such practice speaks to extrapolation. Scientists do not take some test tube compound tested with animals in the laboratory and naively assume that, as a dosage-specific commercial pill, it will be as effective at home as in the laboratory and with all types of humans as opposed to rats. Instead, scientists have developed animal models for extrapolating from data on rats, say, to conclusions about humans and other complex systems; and government agencies have institutionalized multistage procedures, each with its own increasingly general goal, for moving along a long-winded research sequence from discovering a causal relationship to testing its routine application at scale and even monitoring its side effects over even longer periods.

The social science analog to assuming that a replicated relationship will continue to replicate until proven otherwise might be the willingness to infer that one or more independent replications raises the odds of the same relationship being found in still unexamined *UTOSTi contexts. In this scenario, even a single replication changes the relative odds of valid causal extrapolation, even if the absolute odds are still long ones. At its most naïve, this Bayes-like view entails two likely fallacies. The first is that higher relative odds make the costs of incorrect extrapolation tolerable, even though the new and better odds may still be very poor and the costs of incorrect extrapolation might be to institute a policy that at best does not work as planned. The second fallacy is that our consensual standards for describing causal relationships (internal validity) are justifiably stricter than the manifestly laxer standards for generalizing such relationships (external validity). The usual rationale for this is that "internal validity is the sine qua non for external validity" (Campbell & Stanley, 1963), making internal validity logically prior to external validity. While this position is incontrovertible in logic, it fails to recognize that inferences about internal validity are inevitably probabilistic so that, in actual research practice, a trade-off often arises between the resources needed to improve internal and external validity at the margin. The policy sciences would probably do better to examine the potential of natural science practices around extrapolations from animal models and around research stages prior to full scale up as opposed to adopting their practices about the generalization value of one or a few successful replications of a given causal relationship.

A core task of science is to identify *causal mediating processes* that are general in their conditions of application and dependable in their socially important consequences. If we were to know, for example, that engaged time on task routinely increases achievement, then it is possible to determine the set of factors that increase this causal construct and will likely increase achievement therefore. The link to extrapolation is that such determinants of engaged time on task may never have been studied before as causes of achievement, but the causal relationship is so secure that achievement will nonetheless increase with any kind of treatment, person, setting, time, or correlate of achievement that increases engaged time on task. Indeed, one benefit of this perspective is that researchers need not identify all the links from the mediator to the effect; it suffices to show that an antecedent affects the mediator. These are the rationales that Cronbach offers for having the policy sciences devote more resources to identifying general causal mediating processes instead of describing the results of causal agents with less potential for general impact.

Cronbach's advice is not easy to follow. Collectively, policy researchers do not now place a high value on establishing the multiple causes and effects of general mediating processes. Current practice seems to depend more on manipulating policy variants that are already part of current practice or that could be put into practice very soon. The emphasis here is on current policy relevance and immediate use rather than future policy relevance and longer research agendas. Perhaps the current interest in methods for causal mediation will also translate into more studies that are explicitly designed to identify powerful mediators and even, in some studies, to manipulate them directly. Good substantive theories elaborate the links from antecedents to mediators and from mediators to ultimate outcomes; and we generally prefer those theories where the key mediator can be instantiated in several different ways and has important outcomes. In this perspective, Lewin's old adage holds: There is nothing as valuable as a good theory. It is a valid view, but a difficult one to realize in practice.

Within a multivariate statistical framework, probably the best available method for causal extrapolation is *response surface modeling* (Box & Draper, 1987; Rubin, 1992; Tamhane, 2012). It could be very roughly presented as follows. Imagine a single factor with many levels and the opportunity to learn about the function linking the different factor levels to an outcome. Now imagine some kind of a factorial experiment with several factors likely to affect the study outcome and, for the moment, continuous levels on each of these factors. This parametric, factorial structure provides the opportunity to examine how the various factors statistically interact and to estimate the curvature describing how the various factors are jointly related to the outcome. All this is in the service of approximating the unknown true response surface. Multidimensional plots can even help visualize the response function generated by these various factors and levels. Such modeling is used to infer what the response would be at any points defined by joint factor levels. If these are novel points that have never been tested before, extrapolation is involved as described earlier. However, in response-surface modeling, these novel level values can lie between those studied in the past, thus turning what was an issue of extrapolation into the better-supported task of interpolation between values within already examined ranges. If the novel values to which extrapolation is desired lie outside previously studied ranges, then the size of the desired extrapolation can be described as well as the stability of the modeled response function over the area empirically examined.

Response surface modeling is not uncommon in some engineering and medical contexts. It is obviously most feasible in fields where many short-term experiments are possible that permit varying a reasonable number of levels on each of several factors. Studies of this form are rare in social policy, especially more recently. In the past, the New Jersey Negative Income Tax experiment used an incomplete factorial design to vary multiple levels of both the dollar benefit level guaranteed to working poor families and the tax rate they had to pay. But current experimental style favors simpler designs with a single factor and few levels; and it does not look kindly on the modeling needed, especially to describe the curvature of response surfaces. Moreover, categorical (rather than continuous) data complicate matters a little since step functions are now involved and not only curvature. In the near future, we cannot expect to see response surface modeling used for causal generalization purposes in the social sciences.

M*eta-analysis* is better articulated as a model of causal generalization. Its relevance to extrapolation comes from two of its main aspects. One is the opportunity to determine some of the conditions that moderate the size and direction of a causal connection between a labeled T and O. Knowing such causal conditionality can provide clues as to when the treatment will be effective under as yet unexamined conditions. Second and more importantly, meta-analysis can afford the chance to take a relationship that has been frequently and successfully replicated in the past under a heterogeneous set of UTOSTi conditions, and inducing from the stability of results and the heterogeneity of testing circumstances that the same relationship will continue to hold under as yet unexamined *UTOSTi conditions.

Both of these rationales depend on having a large and heterogeneous set of studies. However, the studies entering a meta-analysis are rarely sampled in a formal way from all the studies that might be done on a given topic. Nor are they even formally representative of all those that are already done or even that have met methodological criteria for inclusion. Moreover, constant biases operate from irrelevant forces, such as volunteerism, publication bias, and the confounding of program developers and implementation quality. So, for generalization purposes the strengths of meta-analysis are not due to formal sampling theory, contrary to the heroic assumptions some analysts are eager to make. Instead, the strengths are due to the potential to replicate a cause-effect relationship across heterogeneous UTOSTi circumstances. If the results replicate in adequately powered analyses, they can help rule out some third-variable confounds, identify moderators of the causal relationship, and strengthen inductive leaps to *UTOSTi. However, this last warrant comes only from the common sense notion that a causal relationship that demonstrably holds with many different kinds of persons, in many different settings, at many different times, and with many different variants of the treatment manipulation and of the outcome measure will continue to hold under some as yet unexamined conditions. This last is how meta-analysis is most likely to facilitate causal extrapolation.

A major practical problem for meta-analysis is that (1) many completed studies are required with adequate methodological quality and (2) a wide array of sampling features is needed across each of the five utosti dimensions. Yet, most compendia of effective practices reduce the needed heterogeneity by insisting on inventorying only randomized control trials, perhaps adding a limited range of quasi-experiments. The archives are also often limited to brand-name programs as the treatment under examination; or else they choose narrow topics, such as fourth-grade math curricula. As a result, most inventories contain very few studies on a given topic, reducing the chances of probing true causal heterogeneity. Things will improve if broader causal topics are chosen and as studies accumulate, provided that the funding community is willing to support not just replications in general, but also replications that deliberately vary those UTOSTi attributes that are most likely to condition the direction of a past causal claim. Meta-analysis' potential for useful extrapolation is real, but it is theoretically limited when compared to the potential of response surface modeling. And realizing its potential entails more frequent access to many datasets with multiple sources of heterogeneity across the five domains of generalization captured by utosti.

It is important not to confuse the sample size of studies with heterogeneity to examine the factors most likely to condition the direction of causal impact. In single studies with large sample sizes and broad sampling frames, it is usual to test for causal heterogeneity within obvious limits set by statistical power. Even with thousands of studies on a topic, it is possible to be without important sources of heterogeneity. For instance, there are now thousands of studies on the effectiveness of psychotherapy. Many examine some theory-relevant variant of an established practice that is tested in a university counseling center, and the test is used to fulfill a doctoral requirement. Descriptive detail already exists to map the UTOSTi realm in which psychotherapy practice occurs in the United States today—it is done, for instance, by individuals predominantly between 40 and 60 who, however theoretically trained, have evolved their own eclectic mix of services for improving a given outcome, who treat persons with many different presenting symptoms, who operate alone or in a small group practice, whose offices tend to be in shopping malls, and whose mix of clients, services, and mental health outcomes changes with history, including with reimbursement priorities. Such details describe the UTOSTi to which the sampled studies in a meta-analysis should correspond if it is to be representative, and they also permit researchers to rate past studies according to their level of correspondence to psychotherapy as it is actually practiced. But when this is done (Shadish et al., 2000), the sample of studies potentially available for meta-analysis is shown to be heavily biased toward finely tuned therapy variants from within a particular theory, toward a limited range of outcomes, to university settings, to inexperienced psychotherapists with no current need to gain a living from psychotherapy, and to institutional shelters from historical events that impinge on client flows, paperwork requirements, and reimbursement details. Despite the many psychotherapy effectiveness studies, few reflect generalization targets that describe current practice nationally or that can be used to extrapolate past results to as yet unstudied populations of patients, therapists, settings, times, and treatment variants. Even so, meta-analysis is probably our currently best practical approach to a satisficing methodology of causal extrapolation.

In many fields, from meteorology to macro-economics, attempts are made at causal extrapolation through *causal modeling*. Substantive models are used to estimate what will happen at a later time if one or more predictor variables in the model take on different values. The main variables of interest are those linked to manipulable variables that might be the target for future policy changes, but they can also include different moderator variables from *UTOSTi. Some of the predictors in question may never have been studied in the past, nor in the range currently being examined, or with the most recent updates of some causal links and parameter values. Such causal modeling enjoys less favor now than earlier in both applied microeconomics and most of the social sciences. Many attempts at predicting the consequences of varying some potential policy lever have not been successful, probably due to a combination of incomplete substantive theory, postulated bivariate causal links that have never been tested free of selection bias, proxy measures substituting for missing variables of greater theoretical relevance, and models that do not have a long history of intensive critical scrutiny and readjustment after predictions of changes that did or did not eventuate as expected. The current intellectual climate does not make a resurgence of causal modeling likely in those policy sciences that use social science as their substantive base—macro-economics being the most salient exception. Yet, it is hard to see how extrapolation can be done without some modeling, given the premise that extrapolation requires using available evidence to generalize to *UTOSTi targets that have never been directly studied together. The natural sciences acknowledge this need for some modeling through their partial dependence on, say, animal models to generalize from sample data to inferences about humans.

Scale-up studies involve the penultimate stage in some sequence of ever more policy-realistic studies. The sequence typically begins with discovering a possible causal association and then proceeds to testing it in laboratory and other well-controlled settings. Some version of scale-up comes next that seeks to replicate earlier results with larger samples from a wider range of venues under conditions that best approximate what practice might look like if the treatment were widely adopted and implemented by practitioners and not program developers. Underlying such thinking is the assumption that the target UTOSTi can be clearly defined and that study particulars can be sampled to represent them, at least in terms of category membership if not in terms of formal sampling or construct validation methods. The larger samples also allow empirical probes of whether the same causal relationship is obtained—at least the same causal direction if not effect size—across various subclassifications within the intended UTOSTi domains. Breaking results down by potential moderator variables is already standard practice in individual studies and meta-analyses, but the larger scale-up samples permit probes of the stability of causal direction across multiple, heterogeneous variables deemed likely to condition the direction of the causal relationship. Stable replication across these factors then invites the inductive leap that the same relationship might be expected in the future with as yet unexamined populations of persons, settings, times, and even cause and outcome variants.

CONCLUSION

Whether understood as representation, extrapolation, or both, causal generalization offers daunting challenges. First, there are five not perfectly independent dimensions of generalization in utosti. Second, the optimal methods for generalizing to UTOSTi—random selection and full construct validation—are usually not feasible, and satisficing alternatives have to be chosen based merely on surface similarity and category membership. Third, there are really no optimal models for extrapolating to *UTOSTi, and the best two methods are rarely if ever feasible with social science knowledge—full knowledge of all the antecedents of a general causal mediating process or some variant of response surface modeling. Instead, satisficing methods such as meta-analysis and deliberate sampling for heterogeneity across some aspects of *UTOSTi are all that is pragmatically available and used today. Given this state of affairs and the tyranny of feasible over best practice when it comes to external (but not internal) validity, we are likely to continue with causal generalization processes that reflect the little researchers already know how to do, with little felt need to discover and implement better practices.

The study of external validity has been in the doldrums for a while, and current attempts to advance it are timid, devoted to first steps in moving from u to U. Such steps are better than nothing, but they are not what is needed since the policy sciences will always need to take past findings and extrapolate them to the future or to populations of persons and settings not yet examined or to novel cause and effect variants. Since policy studies depend on causal representation and extrapolation, how can they flourish without an explicit methodology of these two framings of causal generalization? Professional organizations concerned with policy research need to set a more high-profile agenda concerning the importance of causal representation and extrapolation, as they did earlier for internal validity that resulted in the advocacy of random assignment as it has long been practiced in statistics, agriculture, medicine, and psychology. But bivariate causal claims about T and O are inevitably embedded in U, S, and Ti contexts that might condition the T to O relationship; and future targets of generalization—*UTOSTi—will rarely be identical with whatever generalization targets past studies have captured. How are we to proceed with external validity in a way that includes discussion of causal representativeness *and* causal extrapolation?

THOMAS D. COOK is a Professor at IPR, Northwestern University, 2040 Sheridan Road, Evanston, IL 60208. (E-mail: t-cook@northwestern.edu.)

ACKNOWLEDGMENTS

This work was supported in part by the Institute of Education Sciences grant #R305D100033 and National Science Foundation grant #DRL-1228866.

REFERENCES

Box, G. E., & Draper, N. R. (1987). Empirical model-building and response surfaces. Hoboken, NJ: Wiley.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago. IL: Rand McNally.

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

Mackie, J. L. (1974). The cement of the universe: A study of causation. Oxford, UK: Oxford University Press.

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? Journal of Educational and Behavioral Statistics, 17, 363–374.

Shadish, W. R., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. Psychological Bulletin, 126, 512–529.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174, 369–386.

Tamhane, A. C. (2012). Statistical analysis of designed experiments: Theory and applications. Hoboken, NJ: Wiley.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. Journal of Educational and Behavioral Statistics, 38, 239–266.