Journal of MultiDisciplinary Evaluation
Volume 11, Issue 24, 2015

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Using Impact Evaluation Tools to Unpack the Black Box and Learn What Works

Laura R. Peck
Abt Associates Inc.

Background: Researchers and policymakers are increasingly dissatisfied with the "average treatment effect." Not only are they interested in learning about the overall causal effects of policy interventions, but they want to know what specifically it is about the intervention that is responsible for any observed effects.

Purpose: This paper discusses Peck's (2003) approach to creating symmetrically-predicted subgroups for analyzing endogenous features of experimentally evaluated interventions and then it identifies several possible extensions that might help evaluators better understand complex interventions. It aims to enrich evaluation methodologists' toolbox, to improve our ability to analyze "what works" in addressing important questions for policy and program practice.

Setting: Discussion of challenges and possible solutions centers specifically experimentally-designed program evaluations. An illustration comes from a national sectorial training program evaluation.

Intervention: NA

Research Design: The analytic methods examined build on experimentally-designed evaluations and enhance what can be learned from experiments.

Data Collection and Analysis: NA

Findings: After presenting a primer on the Analysis of Symmetrically-Predicted Endogenous Subgroups (ASPES), the paper highlights how some extensions might be especially useful for unpacking the black box, specifically in the face of complex interventions. These extensions include capturing continuously-measured mediators, considering multiple mediators, and considering complexly-measured mediators as well as design options for achieving the same goal.

Keywords: *experimental design; methods; black box.*

Researchers and policymakers are increasingly dissatisfied with learning about only the "average treatment effect." Not only are they interested in learning about the overall causal effects of policy interventions, but they want to know what specifically it is about the intervention that is responsible for any observed effects. While formative evaluation has long been interested in these kinds of questions—explaining why and how an intervention operates, and how these explanations might tie to observed outcomes and impacts—summative evaluations have tended to focus their energies on ensuring that estimated impacts can be interpreted as causal. With this emphasis, summativists come from a design perspective that prioritizes experiments allowing causal attribution of policy impacts. These tend to be coarse tools, identifying whether an intervention had an effect *on average*, probably something about the magnitude of that effect, and possibly something about the extent to which that effect might vary for some subgroups.

This work aims to enrich evaluation methodologists' toolbox, to improve our ability to analyze "what works" in addressing important questions for policy and program practice. It does so by highlighting the current active research and imminent research directions that originate in my 2003 article in the *American Journal of Evaluation*, in particular to inform how the approach established there might be extended to understand the causal effects of complex interventions. In the original work, I established an approach to creating subgroups that can be evaluated experimentally while answering questions about factors that arise non-experimentally. In recent work, colleagues and I have dubbed this an "analysis of symmetrically-predicted endogenous subgroups" (ASPES; see Bell & Peck, 2013; Harvill, Peck & Bell, 2013; Peck, 2013). As my recent article points out, many applications—both direct and indirect—of this analytic approach exist across health (Fernald et al., 2008; Macias et al., 2008), education (Kemple & Snipes, 2000; Schochet & Burghardt, 2007; Unlu, Bozzi et al., 2011; Unlu, Yamaguchi et al., 2011; Zhai et al., 2010; Zhai, Raver & Jones, 2012; Zhai, Raver & Li-Grining, 2011) and social welfare (Gibson, 2007; Harknett, 2006; Morris & Hendra, 2007; Peck, 2005, 2007; Wood, Quinn & Clarkwest, 2011) policy evaluations. Although, as pointed out (Peck, 2013), considerable variability exists in this body of research with respect to its adherence to the spirit and principles espoused in the original (Peck, 2003) work.

Before turning to the ASPES approach, its extensions and future challenges, let me first provide some brief background on motivation for this line of research. Through our intellectual history, scholars have identified shortcomings in the basic analytic approach of treatment-control comparisons, with recognition that "heterogeneity" poses challenges to evaluators. For example, both the heterogeneity of experimental samples and the heterogeneity of treatments offered have garnered attention. Manski pays attention to the "mixing" problem for identifying experimental treatment effects in situations where we mandate homogenous treatments but also find variation in treatments of relevance (e.g., 1995, 1996, 1997, etc.). Other analysts, especially in the medical science, are concerned with treatment compliance and also with dose-response (e.g., Imbens, 2000; Zanutto, Lu & Hornick, 2005). The now classic Angrist, Imbens and Rubin (1996) article considers treatment compliance and uses an instrumental variables (IV) approach to estimate the effect of treatment on those who comply with treatment status. A narrow version of IV is Bloom's (1984) no-show correction. The broader framework within which IV fits is that of principal stratification (e.g., Frangakis & Rubin, 2003), which provides a way to dissect heterogenous treatment effects by establishing "strata" within which principal effects can be estimated. Much of this literature either fuels or at least acknowledges the tension between experimental and non-experimental designs and analytic methods in terms of ability to infer a causal relationship between policy/program and effect. As a result, proposed approaches to estimating the effects of some program or policy hinge on debatable assumptions. For instance, in the context of experimental evaluation research the IV's "exclusion restriction" is necessary to estimate causal effects for treatment compliers, but myriad circumstances exist where it does not hold. The overly restrictive assumptions in this literature inspire and motivate my work and new directions regarding "complexity" addressed in this paper.

The paper proceeds as follows: First, I provide a simple introduction to the ASPES approach, providing illustrative examples. Then, discussion proceeds both (1) to identify elements of the analytic approach where current research is exploring refinements and extensions that intend to make the approach both better understood among evaluators and more flexible to varied applications, and (2) to identify future areas for research that seem promising to help unpack the black box and learn what works, especially within a context of complex interventions.

An Introduction to Analyzing the Effects of Policy Interventions on Endogenous Factors: A Primer on Peck (2003)

The ASPES approach uses observed baseline characteristics, which are exogenous to the treatment indicator, to create experimentally valid subgroups that predict some post-randomization characteristic of the sample. This characteristic can be observed in the treatment arm—such a compliance with treatment, experience of high treatment dosage or quality, or experience of some element of a multi-faceted intervention—or it can be observed in the control arm—such as risk of school dropout or earnings levels in the absence of the intervention. The analysis involves the following three stages: (1) predicting subgroup membership; (2) estimating predicted subgroup impacts; and (3) converting impacts from representing members who are predicted to be in a specific subgroup to represent those members who are actually in a specific subgroup. The two key features that distinguish the ASPES approach from other, related work are (a) assuring symmetric identification of endogenous subgroups,[1] and (b) converting impacts from representing predicted subgroups to representing actual subgroups. These two features are about internal and external validity, respectively. All three analytic stages are detailed next.

Stage 1

At first, the analyst must identify subgroup membership, using a strategy that ensures symmetric prediction. In Stage 1, the probability (or propensity) of being in a subgroup (e.g., high dosage subgroup) is modeled as a function of baseline traits, using a logit, probit or linear probability model. Either a split sample or cross-validation approach will ensure that the subgroups are symmetrically-identified, such that no one is any better identified than the other, thereby retaining the integrity of the experimental design.[2] The aim of this prediction process is to identify which treatment and control group members have the baseline characteristics that associated them with being the in subgroup of interest (e.g., being exposed to high (or low) program dosage, etc.).

The result of this first-stage prediction is a continuous score that represents subgroup membership. The score can be used as is; and this "continuous" version of ASPES was established in Peck (2003) and is a current topic of methodological inquiry. Most commonly, however, in applications and extensions of the approach is slicing the continuous score into discrete subgroups. For example, subgroup members can be defined as "high dosage" if their predicted probability of being in this subgroup is greater than or equal to 0.50; otherwise, they would be in the low-dosage subgroup. Examining the distribution of scores may help assess where the logical breakpoints are, and that 0.50 may be the optimal breakpoint. Another possibility is to select the breakpoint that maximizes correct placement of individuals into the correct subgroups of interest. Other applications of the approach have used quartile (25th and 75th percentiles) to designate the highest and lowest risk subgroups (Moulton, Peck & Bell, 2014).

Stage 2

With treatment and control subgroups identified, Stage 2 involves estimating subgroup impacts on those predicted subgroups. Because the subgroups are identified as a function of baseline (exogenous) characteristics, the impacts by subgroup are the same as any other experimentally-defined subgroup. In this case, however, the subgroup—rather than being a single trait such as education level or sex—is a composite of traits that is associated with some post-randomization choice, event, pathway or experience. As such, the mean difference between treatment and control subgroups is an unbiased estimate or the program effect for that subgroup. A simple split-sample subgroup analysis can be undertaken to generate

[1] By "symmetric" I mean that the treatment and control subsets are equivalent in all ways, both measureable and unmeasurable, as would be expected of any subgroup that would be part of an experimental subgroup analysis. In other words, neither the treatment nor the control subgroup is better or worse identified than the other.

[2] Using the entire treatment group for subgroup prediction and for impact analysis can introduce bias because of the better fit that is inevitable for the sample that is used for modeling (e.g., Harvill, Peck & Bell, 2013; Peck 2003).The cross-validation approach ensures symmetric prediction while retaining the entire sample for the analysis stage (Harvill, Peck & Bell, 2013) and is preferable to a full jackknife approach (Abadie, Chingos & West, 2014).

these estimated impacts for two subgroups, call them L and H, for low and high dosage, as follows:

$$I_H = \bar{Y}_{TH} - \bar{Y}_{CH}$$
$$I_L = \bar{Y}_{TL} - \bar{Y}_{CL}$$

Impacts can be estimated, by subgroup, using standard multivariate regression to increase precision by controlling for random baseline variability, as follows:

$$y_i = \alpha + \delta T_i + \beta X_i + \varepsilon_i$$

where,

y is the outcome;
α is the intercept (interpreted as the control mean outcome);
T is the treatment indicator (treatment = 1; control = 0);
δ is the impact of the treatment;
X is a vector of baseline characteristics;
β are the coefficients on the baseline characteristics;
e is the residual; and
the subscript i indexes individuals.

In the above regression equation, $\hat{\delta}$ is the estimate of the impact of the treatment on the predicted subgroup of interest. For example, when the sample is limited to the set of members who are predicted to be in Subgroup H, $\hat{\delta}$ is an estimate of the impact of the intervention on members who were predicted to be in Subgroup H (I_H from the simple difference-of-means equations above).

This impact estimate is free from bias from selection and other sources: it is an unbiased estimate of the impact for the predicted subgroup. That said, the predicted subgroup contains a blend of individuals who are actually in that subgroup and individuals who are not, misplaced there by virtue of imperfect prediction in Stage 1. In other words, while this impact estimate is internally valid, its external validity may be limited, which is what motivates the third analytic stage in the ASPES approach.

Stage 3

Stage 3 involves converting the impacts from representing the predicted subgroups to representing the actual subgroups, by assumption. As noted, the results from the Stage 2 analysis are purely experimental. This means that the Stage 3 conversion rests on an experimental foundation. There is no bias (from selection or other sources) in the impact estimates because of the experimental design, and so the extent to which one believes the results from the conversion can also be interpreted as experimental then rests on comfort with the assumptions needed to make the conversion.

As above, assume two subgroups: Subgroup H and Subgroup L. I have posited that one can think of the estimated impacts on each of the two predicted subgroups (H and L) as a weighted sum of the impacts on those who are actually in that subgroup and those who are actually in the alternative subgroup. For instance, the following equation states that the impact on predicted Subgroup H members is a weighted sum of the impacts on actual Subgroup H members and actual Subgroup L members, where the weights represent the proportion of predicted Subgroup H members who are actually in Subgroups H and L, respectively:

$$I_H = w_H H_H + (1 - w_H) L_H$$

where

I_H is the impact on predicted Subgroup H members;
w_H is the proportion of predicted Subgroup H members who are actually in Subgroup H;
H_H is the impact on predicted Subgroup H members who are actual Subgroup H members; and
L_H is the impact on predicted Subgroup H members who are actual Subgroup L members.

Similarly, the impacts on Subgroup L can be thought of as a blend of the impacts estimated for those who are predicted to be and who are actual subgroup members, with the weights being the proportion of those predicted as such who are actually in each subgroup, as follows:

$$I_L = w_L L_L + (1 - w_L) H_L$$

where

I_L is the impact on predicted Subgroup L members;
w_L is the proportion of predicted Subgroup L members who are actually in Subgroup L;
L_L is the impact on predicted Subgroup L members who are actual Subgroup L members; and

H_L is the impact on predicted Subgroup L members who are actual Subgroup H members.

In practice, w_H and w_L are only directly observable for treatment group members and are unknown for control group members. We therefore calculate the values of w_H and w_L from treatment group data and assume that w_H and w_L are the same for treatment and comparison group members. We can deem this as a safe assumption to the extent that we believe that the split-sample or cross-validation approaches to identifying those subgroups result in an equally good fit of the model for both treatment and comparison group members (they are designed to do so and have been shown to be effective as such in follow-up research, shown in Harvill, Peck & Bell (2013) and Abadie, Chingos & West (2014)).

Together, these equations contain four unknowns, and so some additional assumptions are necessary in order to solve the system.[3] As I suggested in Peck (2003), in the discrete two-group case, a homogeneity assumption will meet this requirement, as follows:[4]

$$H_H = H_L$$
$$L_H = L_L$$

These two equations state that, regardless of which subgroup the actual subgroup members are predicted to be in, the impact on them is the same on average. This means that for individuals who are actually in Subgroup H, the impact of the intervention is the same regardless of whether the participant is predicted to be in Subgroup H or Subgroup L. With this homogeneity assumption, the system of two equations can be rearranged to solve for the unknown elements of interest as a function of the known elements computed in Stage 1 as follows:

$$H = \frac{(w_L)(I_H) - (1 - w_H)(I_L)}{w_H + w_L - 1}$$

$$L = \frac{(w_H)(I_L) - (1 - w_L)(I_H)}{w_H + w_L - 1}$$

If the assumptions embedded here hold, then the estimators based on these final two "conversion" equations are asymptotically unbiased.

As described here, the discrete ASPES method is a useful avenue for estimating program impacts for subgroups defined by some post-randomization mediator that can be modeled as a function of baseline variables. While the foundation exists for retaining the continuous score and analyzing its impact accordingly, I include this option in my discussion of extensions, next, because it supports specifically the exploration of more complex mediators, ones that are not readily discretized/simplified.

Research on Elements and Extensions of the ASPES Approach for Better Understanding Causality and Complexity

Before discussing the specific research activity and analytic extensions, this section first describes the problems that compel these investigations.

Motivating Problems of Complexity

The ASPES approach was developed initially as an alternative to imposing a no-show assumption. In that situation, the problem of interest pertained to those who did not take up the opportunity extended by treatment but could not be assumed to have experienced no effect because of treatment activities that targeted non-takers (Peck, 1999). Whether individuals take up the offer of treatment is a simple problem, and the problems that plague today's evaluations are much more complex.

To elaborate, many major, current national evaluations involve multi-faceted treatments, where individuals' participation in certain facets of the treatment is of relevance to policy and practice. These evaluations are interested in which of these program facets are drivers of the program's overall effects, knowledge of which can then be used to design the most efficacious programs possible. This relates to what I will call "programmatic complexity," which I will describe next. I then address another source of complexity—"temporal

[3] Another non-controversial assumption is implied: that the distribution of predicted subgroup membership is the same, relative to the actual subgroup membership, between treatment and control groups. This is untestable because actual subgroup membership is known only in one experimental arm.

[4] A three- or more group analysis can also use a comparable homogeneity assumption, although more alternatives exist there than do in the two-group case. As Moulton, Peck and Bell (2013) explore, a no-show assumption can be coupled with a partial homogeneity assumption; and as Bell and Peck (2013) elaborate, other plausible assumptions exist as well.

complexity"—which also poses challenges to standard evaluation practice. Other evaluation scholars have considered the concept of "complexity," defining it in contrast to interventions that are "simple" or "complicated" (e.g., Patton, 2010). The metaphors are that "following a recipe" is "simple," "sending a rocket into space" is "complicated," and "raising a child" is "complex" (Patton, 2010, p.92). The programmatic and temporal dimensions of complexity I discuss next fit within the evaluation field's commonly held definition of complexity.

Programmatic complexity. Particularly in the social policy arena, programs are complex, offering varied combinations of features, often together in a package that diverse administrative and management structures are involved in delivering through varying implementation practices. For example, the U.S. Department of Health and Human Services, Administration for Children and Families' Health Profession Opportunity Grants (HPOG) program funds grantees who offer low-income individuals a multi-faceted career pathways-based education, training and services targeting occupations in the healthcare field. HPOG program grantees vary in their administrative configuration: some grantees are community colleges, others workforce investment boards; some are centralized while others decentralize delivery to a network of partner organizations and nonprofits. This is one layer of variation that is an umbrella to both the implementation practices and the actual program content. Implementation practices vary for a variety of reasons, related to administrative and management structure, programmatic content and offerings, and target population characteristics and needs. HPOG programs offer various combinations of intake and assessment services, academic advising and career counseling, support services and financial and non-financial assistance. While all of the programs fall under the general rubric of being designed as "career pathways" programs, the theory itself implies complexity: by default, customization implies that programs will and must vary not only across the places that follow the same program model but within locations, depending on the specific needs of program enrollees. Finally, program enrollees and targets come with their own needs and desires, some of which are anticipated in program design and implementation, and some of which are idiosyncratic and inspire additional program customization.

As a result, we might classify the dimensions of complexity into the following: theory, administration, implementation, program design, and targets and participants. Consider the analogy of a cafeteria: the program theory is the recipe that chefs—funders and administrators—use in deciding what to put on the buffet. There are many options on the buffet that local managers will choose from in order to fill a tray with the right items to fit the nutritional needs of their program participants and targets. Those managers then serve the meal to their participants, who may or may not eat everything on the plate. If one considers this an elementary, middle or high school cafeteria, then there is an additional level of complexity (or chaos) in the negotiation of items that people might bring from home. Figure 1 presents visually this analogy.

The exhibit shows the following: A is the Kitchen (program design), where funding, administrative and management decisions take place; B is the Buffet (program offerings), where the results of decisions about what to offer are evident; C is a Server (local management), who decides what to bring to participants and how to deliver it; D are Diners (participants), consuming what is brought to them, including some of whom supplement with outside food; and E is the Queue (additional program managers or targets), getting in line to access and deliver services.

Each iteration of a program or policy in practice must differ: a grantee in Pennsylvania cannot, even if it wanted to and tried, exactly replicate an "HPOG program" as implemented in Connecticut, for example. To illustrate, both Central Susquehanna Intermediate Unit (Pennsylvania) and The WorkPlace (Connecticut) grantees serve several counties, require participants to complete soft skills trainings, and provide one-one-one case management as well as child care and transportation assistance if participants are unable to get them from other sources. However these two grantees serve very different populations: Central Susquehanna Intermediate Unit serves mostly white participants within a large, generally rural area in Pennsylvania, while The WorkPlace serves more urban counties in Connecticut and over half of its participants are black or African American. The WorkPlace also serves almost twice as many participants who received TANF benefits at intake than Central Susquehanna Intermediate Unit. Moreover, the two grantees use different approaches for providing similar services to meet the needs of their specific target populations. Central Susquehanna Intermediate Unit HPOG participants meet the soft skills training requirement by attending a minimum number workshops offered by HPOG staff or partners in

varied locations and some are even offered on DVDs so that transportation will not be a barrier to participation. While The WorkPlace requires HPOG participants to attend a week long in-person soft skills workshop, this is offered at only one location.

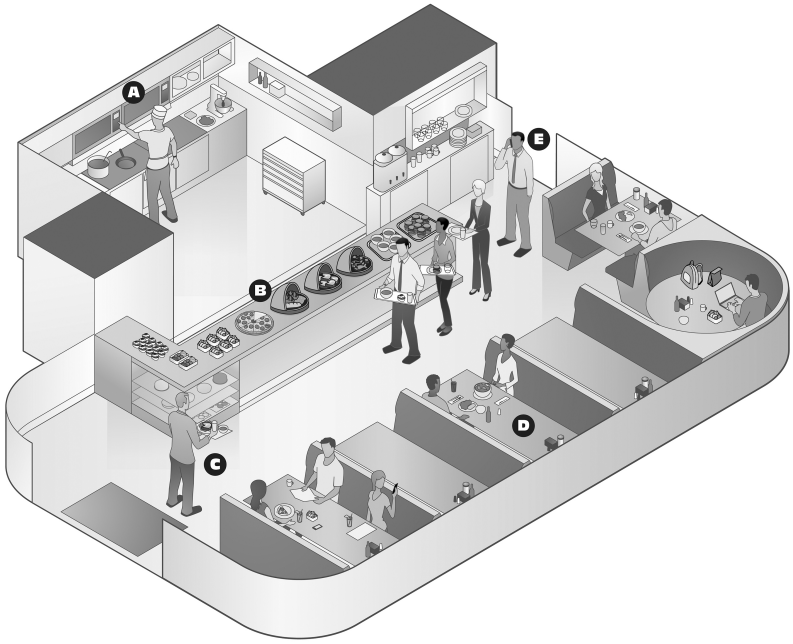

Exhibit 1. Illustration of Cafeteria as Program Complexity

Now consider 32 rather than two grantees, and this complexity challenge becomes greater still. Even if exact replicates could be made across many locations, at the very least programs aim to do lots at once, packaging various elements of a vision for policy change into one so-called program. No longer are we talking about experiments as simple as the National Income Tax (NIT; what percentage tax rate induces what behavioral response, with implications for subsequent earnings); but instead we want to know how putting several incentives (and punishments) together with varied supports to several training and educational offerings might collectively induce some change. At the same time that we want to know how everything works together, we are interested in pulling it apart. That is, evaluations are often charged with reporting the program's overall treatment effect, but they also increasingly want to report about the relative effects of the individual component parts.

Temporal complexity. This complexity is compounded when we consider time as an added dimension: programs are not only implemented and evaluated at different points in time but they also evolve over time, in ways that might interact with other traits, including programmatic content/design as well as implementation features and surrounding contexts. One program—consider HPOG again—may have its own over-time trajectory, and that becomes multiple trajectories when we consider the many places operating the programs. Variation in calendar time is also associated with many variants of seasonality, including the economic cycle, political regimes, school years, fiscal years, population shifts, and general secular changes. Trying to capture the influence of any one of these dimensions poses evaluation challenges. It is a main reason that experimentally-designed evaluations are used: to net out the effect of "historical" threats to internal validity, at least eliminating contextual/temporal influences on participants' changing outcomes. Even so, within a fixed time period, both treatment and control conditions may change; and a standard experimental design will capture the effects, averaged over that period, obscuring anything about the temporal variation, some of which might be particularly interesting, important or relevant to policy-making or program re-design.

The analogy portrayed in Exhibit 1, therefore, might be considered a programmatic snapshot, a cross-section of the programmatic complexity that exists. Adding the time dimension would transform the exhibit into a video (motion picture), a longitudinal portrayal of the goings-on in the cafeteria. How to know what impacts the

program is having in this situation is practically mind-boggling: "the program" itself varies in *what it is* not only across many places but also over time, with the location-time interaction also varying. Conventional evaluations—whether formative or summative, and experimental or non-experimental in their design—tend to simplify these complexities, generally out of self-preservation. I would argue that only a formal Theory of Change approach to evaluation comes close to capturing these dimensions of programmatic and temporal complexity; and even that falls short, again usually out of the necessity to do so. In social science research—particularly evaluations that cover many people and places—we tend to follow nomothetic models of causality, reserving the ideographic models for ethnographic and psychological case study sorts of work. But, as we move toward answering the kinds of questions that are more flavorful than the vanilla average treatment effect, we get pushed into a realm, where we must face the realities of the world's complexities, rather than simplifying them away.

Possible Responses

This section explores how extensions of the ASPES approach may offer insight into these kinds of complexities. To start, the basic ASPES approach is designed to address questions of what works, and it does so effectively and without bias in the circumstance where one has baseline data that can usefully model some post-treatment choice, experience or mediator of some sort. To date, most related research has focused on whether one is "in" or "out" of some such group, although increasingly the approach is being extended to multi-group cases (e.g., Bell & Peck, 2013; Peck & Bell, 2014). The future research directions and opportunities discussed include: using continuous mediator measurement (rather than discrete, as described above); moving from single to multiple mediators; considering complex groupings of program experiences that can tell us more about varied "packages" of social policy assistance; and using design options—such as multi-arm, factorial and mega-multi-arm trials—to narrow in on the impacts drivers in complex interventions.

Continuous Mediators. As noted above, applications and extensions of the approach have focused on the discrete version, but having the full information that comes from a continuous mediator variable may be more valuable. Substantial opportunity exists to extend the ASPES method to continuous mediators, allowing researchers to gain a more nuanced understanding of the relationship between mediators and program impacts.[5] In this situation, prior work (Peck & Moulton, 2013) has proposed the following equation to estimate the impact of the continuous predicted value of the mediator M on an outcome Y:

$$Y_i = \beta_0 + \beta_1 \widehat{M} i + \beta_2 T_i + \beta_3 \widehat{M}_i T_i + \varepsilon_{2i}$$

where,

Y is the outcome being examined; $\widehat{M}$ is the predicted value of M generated from Stage 1; T is the treatment indicator (treatment = 1; control = 0); and ε_2 is an error term that captures all other factors that influence the outcome.

Then, the impact of the treatment on outcome Y is given by:

$$\frac{\partial Y}{\partial T} = \beta_2 + \beta_3 \widehat{M} i$$

This model assumes a linear impact curve of $\widehat{M}$ with intercept β_2 and slope β_3.

The intent of current research is to establish the assumptions necessary to interpret the effect of the predicted mediator variable as representative of the actual mediating factor of interest. One possibility is a simple extension of the homogeneity assumption from the discrete to continuous case: where the impacts are assumed to be the same, on average, for those with a given value of M and those with a given value of $\widehat{M}$. In other words, the necessary assumption for drawing conclusions about sample members with an actual value of M when we base the analysis on sample members with a predicted value, $\widehat{M}$, is that the impact on those with an actual value of M is the same, on average, as the impact on those with the predicted value, $\widehat{M}$ (Peck & Moulton, 2013). Early discussions with fellow methodologists reveal that some alternative assumptions might be relevant here as well, including two things: (1) a version of the exclusion restriction commonly used in instrumental variables estimation; and (2) an assumption about the non-interactive effects of

[5] Recent work by Bein (2013) reflects on this continuous specification, frames it within a potential outcomes framework and likens the estimator to a principal effect. This innovation, therefore, may simply be a recasting of related/prior work for addressing new questions.

competing mediating factors. The exclusion restriction assumes that the sole pathway for the mediator to have its effect is through the instrument, which in this case would be the mediator as predicted by baseline characteristics. The second assumption is that the mediator does not interact—have synergistic, either favorably or diminishing effects—with some other program component that is not being predicted but may still be part of the intervention.

Multiple Mediators. This latter assumption (non-interacted effects of competing mediating factors) and dissatisfaction with it compels additional research that might further develop the methodological toolbox to allow for considering the independent and interactive effects of multiple mediators simultaneously. That is, when multiple mediators can play a role in determining the magnitude of an intervention's impact, basic discrete or continuous analysis of symmetrically predicted subgroups to inform "what works"—just like analyses of demographic subgroups taken one characteristic at a time—risks misattribution of the true causal influence. For example, successfully determining that impacts are larger for individuals with high rather than low treatment dosages using ASPES might lead to the erroneous conclusion that "dosage matters," when in fact what matters is intervention quality and (a) quality is not modeled simultaneously with dosage but (b) correlates positively with dosage. Dosage then becomes a proxy for quality, and a policy change that increases dosage but leaves quality unchanged will not produce a larger impact. This is the standard "omitted confounder" bias problem common to all types of multivariate analyses. It arises regarding non-simultaneous analysis of dosage and quality, or of multiple dimensions of quality when one dimension matters to impact but the other (which correlates with the first) does not, or in many other situations of multiple possible mediational pathways to impact.

The solution, therefore, is simultaneous analysis of two or more potential mediational influences, using the strategy of ASPES or another means. In the ASPES context, this requires defining endogenous subgroups on the basis of the interaction of two dimensions of post-random assignment experience in the treatment group, such as the level of dosage and quality experienced by individual sample members. In that instance, one suggestion is to predict membership of treatment group members in one of four endogenous subgroups—large dosage/high quality, large dosage/low quality, small dosage/high quality, small dosage/low quality—where that classification is observed. Then subsets of the sample can be symmetrically-identified on that basis. The resulting impacts (i.e., the treatment-control group difference in average outcomes) for the predicted subgroups have full internal validity based on their symmetric derivation from an experimentally-divided sample.

The external meaning of these results, however, depends on their relationship to impacts on actual endogenous subgroups—large dosage/high quality, large dosage/low quality, small dosage/high quality, small dosage/low quality. As always with ASPES, translation of impact evidence with internal validity on predicted endogenous subgroups to estimates of impact with external meaning for actual endogenous subgroups depends on the conversion assumptions adopted. As with the earlier one-dimensional case, this translation requires that findings are robust to the conversion assumptions. Work underway with colleagues considers a variety of conversion assumptions ranging from the homogeneity assumption—that impacts on participants with a particular dosage and quality combination are the same regardless of the subgroups to which sample members are predicted to belong—to assumptions concerning the additivity (versus super-additivity, when synergisms occur) of impact magnitudes when larger dosage is added to higher quality or vice versa (see Bell (2013); and the Appendix to Peck & Bell (2014) offers an example).

Complexly-measured Mediators. In addition to continuously-measured mediators and multiple mediators, this notion of "complexly-measured" mediators warrants elaboration in order to enrich the evaluator's toolbox for dealing with programmatic and temporal complexity. Earlier work highlights the promise of atheoretical analytic approaches, such as latent class or cluster analysis for identifying subgroups that are more complex than a single indicator of program participation, for example. In a nonexperimental context, Yoshikawa et al. (2001) use a cluster analytic approach to identify service-related groupings of individuals exposed to a welfare reform intervention. Bringing this idea into an experimental context, Gibson (2003) and Peck (2005) use cluster analysis to identify treatment *and* control group subsets that have particular program-related features, with my work focusing on take-up alone and Gibson's considering multiple program characteristics simultaneously. Whether the identifiable types or profiles are meaningful will determine the extent to which this kind of approach is useful for informing policy-relevant questions. When Gibson identified several

program features, and combinations of features, that characterized certain groups of individuals' program experiences, doing so obscured the relative effects of individual program components. As a result, this limits usefulness for program design because it is difficult to parse out the heterogeneity captured in these diverse groups. The clustering approach, at least to date in its two identified applications, is messier than the discrete or continuous ASPES approach, thereby limiting its use for program design. In turn, future applications might consider how to focus more clearly on *specific* combinations of program elements that might better inform policy decisions and program design. Doing so seems a tall order, but should it be successful certainly holds promise to increase how informative results from such an analysis might be.

But also, for this approach to be useful, individual-level baseline characteristics must be able to predict program experience. They might be able to in some circumstances, though may not be in other (perhaps most) circumstances. That is, some of a person's own characteristics—such as her race or marital status—are unlikely to predict well what program features she accesses. Other characteristics—such as education and parental status—might be more predictive: if level of education tracks to how she engages with the program and number or ages of children predict use of child care supports, then program experiences might be effectively predicted using these kinds of variables.

What either of these types of variables cannot do is predict how programs might *change* over time, in response to needs to understand temporal complexity. At the very least, in the context of a multi-site evaluation, a site-level indicator might be a reasonable proxy for those site traits that would be associated with changes in particular program offerings. Future research might explore the extent to which site indicators or other site-level variables might be useful in predicting program change and trajectories such that those experiences can be captured and their effects analyzed.

What might be more promising still is an extension of this to cluster-randomized experiments, where the unit randomized—a school, organization or other aggregate unit—is the unit where decisions about program change are made. If baseline traits of these sites can predict future programmatic directions, then the ASPES approach—by itself or in conjunction with a cluster-analytic approach to predicting/identifying endogenous subsets of an experimental sample—will provide a solid foundation for this kind of extension.

Design Options. As all impact evaluators would agree: design trumps analysis! While the analytic approach and extensions discussed in this paper all *use* the experimental design and therefore are preferable to purely nonexperimental analyses, there are elements that move beyond it and require reliance on assumptions to draw causal conclusions. In response, the discussion I turn to next identifies some ways in which we can use experimental design options for improving our ability to examine complex interventions and tease out answers to varied "what works" questions.

If we can anticipate, in advance, which elements of a multi-faceted treatment we would want to say something about independently, then we should randomize to them. Returning to the HPOG example, the evaluation identified three program components that are of particular interest to policy and practice that warrant especially rigorously-tested answers to the question of the extent to which they are effective ingredients within the HPOG recipe. In several locations, therefore, program staff randomized program entrants to one of three groups:

- a basic treatment group (A), where individuals gain access to the regular[6] HPOG program;
- an enhanced treatment group (A+B), where individuals gain access to the regular HPOG program *plus* a selected enhancement; and
- a control group (C), where individuals do not have access to HPOG but instead can access whatever other services are available to them in their community.

This three-arm design allows testing the added effect of some specific program element. In contrast, a three-arm design such as that employed in the National Evaluation of Welfare to Work Strategies (NEWWS) can test the relative effectiveness of competing program designs. While

[6] By "regular" I mean the program that grantees designed and implemented by choice. The program, then, is not the same everywhere but instead—and in sync with the cafeteria analogy—each is the grantee's selection of program components, implemented in the manner that they choose. Harvill, Moulton and Peck (forthcoming) explores how to make sense of variation in the basic program to learn about varied program components, whether they are offered as experimentally-allocated enhancements or not.

those NEWWS tests were useful in informing questions about *which program model* is more effective, it did nothing to identify *which elements* of the interventions were stronger contributors to the policy's effectiveness. Instead, the kind of three-arm design that HPOG employs is poised do so: it involves a test of treatment A versus treatment A+B, where B is a specific program ingredient. The design allows estimating the effect of the program alone (and without the added ingredient) relative to the enhanced program. With three selected ingredients being tested, the evaluation will be in a position to add to the evidence base regarding which of these three specific program ingredients are worth incorporating within these kinds of multi-faceted training programs.

This approach seems easily replicable and salable to programs on the ground: for any programs considering trying something new or adding new program features, a reasonable way to do so is incrementally, using experimentation along the way with phase-in, in order to determine the extent to which program changes are worthwhile improvements. Although any given test will most likely include samples too small to reliably detect the small incremental effects of an incremental policy change, the amassing of many such experiments can and should add to our collective knowledge.

Taking this to the extreme might involve running many (probably small) experiments where each subgroup of interest is randomized to each program element. While this would certainly narrow the programmatic complexity to units where we might learn about "what works," it would come at a cost of creating substantial complexity for the practice of program evaluation. I will label this a "mega-multi-arm" strategy, one where we envision experimenting in many settings and times, using experimentation to test any chance we might want to try. Any given trial might not yield policy useful results, but the collection of trials that are part of this "strategy" together can, over time, create the evidence base for answering important "what works" questions that respond to both the programmatic and temporal complexity problems described earlier.

While a three- or more-arm design will accomplish learning through experimental tests of given program features, a factorial design will add cross-factor interactions to the list of things we can estimate impacts on. To a certain extent, the factorial design (A, B, A+B, C) can be an efficient use of a research sample; it does so in support of more questions than might be essential. We gain, at the cost of less precision on some treatment contrasts/comparisons, the ability to answer more questions about mixed conditionality (the relative marginal effects of two factors). If untreated samples can be pooled, then sample size is increased; but the contrast "to what" becomes murky because of the blend of true control cases and unexposed treatment cases that fall into that pool. I have concluded that the factorial design, therefore, is not as good for understanding program complexity as "flat" multi-arm designs for this reason.

Conclusion

As this paper has argued, issues of programmatic and temporal complexity create challenges for evaluators. The questions we want answered about such complex programs are complicated as well: while we are interested in the overall effect of richly-configured programs, we are also interested in how specific ingredients contribute to the overall recipe. Designed to respond to basic "what works" questions, my (Peck, 2003) approach provides a solid underpinning for some future extensions that can respond to needs for information about social programs in a more complex world. I have posited that using a continuous measure of program mediators is one possibility that overcomes the sometimes-awkward decision of making a subgroup indicator discrete: when two or three groups are not easily sliced from across a predicted continuous mediator, retaining the continuous score may allow retaining useful information that can inform the effects of that mediator on the outcome. Moreover, methods in addition to regression-based prediction—such as cluster analysis or latent class analysis—hold promise for identifying more complexly-measured mediators, where those mediators are composites of various program features. To the extent that such specific groupings can be identified through these methods, the ensuing analysis could be useful in informing "what works" questions. Extending ASPES to accommodate multiple mediators is another useful direction to pursue. But, as I argue, innovating and expanding in evaluation design is my preferred approach for evaluating complexity: if we could integrate experimentation into common program practice, then we would be in a position to learn continually and, over time, amass sufficient evidence to be confident about which elements of multi-faceted programs are the ones that should be replicated as opposed to the ones that need retooling or retiring.

I acknowledge that I come to this discussion with a bias: I strongly prefer randomized experiments for their ability to provide information about the causal effects of the policies, programs and interventions we run. Colleagues across this evaluation aisle will emphasize that non-experimental methods might be more appropriate for some kinds of questions, particularly those having to do with "complexity" where it seems a stretch to experiment. While I firmly concur that varied questions demand varied methods, questions about *causality* imply experimentation. As a result, this paper has explored how to extend experimental methods for understanding causality and complexity.

Even colleagues on my side of the evaluation aisle will criticize that only the first-stage analytic results—those representing the effects on *predicted* subgroups—are free from bias and that the results of interest (on *actual* subgroups) derive through assumptions, which may or may not be reasonable. Again, while I recognize this criticism, I would respond that—until other methodologists come up with any better options—this is the best we've got. That is, the past decade has seen a clear increase in demand for answers to what works questions, with the ASPES approach, established in my prior (Peck, 2002, 2003) work, being the one approach—certainly among many options (see Peck, 2013)—that *uses* the experimental design and *retains* the experimental treatment-control contrast to inform those answers. As such, it provides the ideal springboard for extensions that make the approach more flexible to more situations, including those about complexity discussed herein. In the end, this paper provides just the beginning of a roadmap for exciting innovations that can enhance the tools that evaluator use to explore "what works" so that policymakers and practitioners can base their future directions on solid evidence.

References

Abadie, Alberto, Matthew M. Chingos & Martin.R. West. (2014). Endogenous Stratification in Randomized Experiments. Cambridge, MA: Harvard University Working Paper. Available at http://www.ksg.harvard.edu/fs/aabadie/stratification.pdf

Angrist, Joshua D., Guido W. Imbens, & Donald B. Rubin. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association,* 91(434), 444-455. DOI: 10.2307/2291629

Bein, Edward. (2013). "Proxy Variable and Other Estimators of Principal Effects " to be presented at the Annual Fall Research Conference of the Association for Public Policy Analysis and Management, Washington, DC, November 7.

Bell, Stephen H. (2013). "Extending Analysis of Symmetrically Predicted Endogenous Subgroups to Multiple Mediators" to be presented at the Annual Fall Research Conference of the Association for Public Policy Analysis and Management, Washington, DC, November 7.

Bell, Stephen H., & Laura R. Peck. (2013). Using Symmetric Predication of Endogenous Subgroups for Causal Inferences about Program Effects under Robust Assumptions: Part Two of a Method Note in Three Parts. *American Journal of Evaluation,* 34(3), 413-426. DOI: 10.1177/1098214013489338

Bloom, Howard S. (1984). Accounting for No-shows in Experimental Evaluation Designs. *Evaluation Review,* 8(2), 225-246. DOI: 10.1177/0193841X8400800205

Frangakis, Constantin E., & Donald B. Rubin. (2002). Principal Stratification in Causal Inference. *Biometrics*, 58(1), 21-29.

Fernald, Lia C. H., Rita Hamad, Dean Karlan, Emily J. Ozer, & Jonathan Zinman. (2008). Small Individual Loans and Mental Health: A Randomized Controlled Trial Among South African Adults. *BMC Public Health*, 8, Article 409. DOI: 10.1186/1471-2458-8-409

Gibson, Christina M. (2003). Privileging the Participant: The Importance of Subgroup Analysis in Social Welfare Evaluations. *American Journal of Evaluation*, 24(4), 443-469. DOI: 10.1177/109821400302400403

Harknett, Kristen. (2006). Estimating Effects for Program Participants Using Propensity Score Does Receiving an Earnings Supplement Affect Union Formation? *Evaluation Review*, 30(6), 741-778. DOI: 10.1177/0193841X06293411

Harvill, Eleanor L., Shawn Moulton & Laura R. Peck. (forthcoming). Health Profession Opportunity Grants Impact Study Technical Supplement to the Evaluation Design Report: Impact Analysis Plan. OPRE Report # XXX, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Harvill, Eleanor L., Laura R. Peck, & Stephen H. Bell. (2013). On Overfitting in Analysis of Symmetrically Predicted Endogenous Subgroups from Randomized Experimental

Samples: Part Three of a Method Note in Three Parts. *American Journal of Evaluation*, 34(4). DOI: 10.1177/1098214013503201

Imbens, Giudo W. (2000). The Role of the Propensity Score in Estimating Dose-response Functions. *Biometrika*, 87(3), 706-710. DOI: 10.1093/biomet/87.3.706

Kemple, James J., & Jason C. Snipes. (2000). Career Academies: Impacts on Students' Engagement and Performance in High School. New York, NY: Manpower Demonstration Research Corporation.

Macias, Cathaleene, Danson R. Jones, William A. Hargreaves, Qi Wang, Charles F. Rodican, Paul J. Barreira, & Paul B. Gold. (2008). When Programs Benefit Some People More than Others: Tests of Differential Service Effectiveness. *Administration and Policy in Mental Health and Mental Health Services Research*, 35(4), 283-294. DOI: 10.1007/s10488-008-0174-y

Manski, Charles F. (1995). Identification Problems in the Social Sciences. Cambridge, MA: Harvard University Press.

Manski, Charles F. (1996). Learning about Treatment Effects from Experiments with Random Assignment of Treatments. *Journal of Human Resources*, 31, 709-733.

Manski, Charles F. (1997). The Mixing Problem in Program Evaluation. *Review of Economic Studies*, 64(4), 537-553. DOI: 10.2307/2971730

Morris, Pamela A., & Richard Hendra. (2009). Losing the Safety Net: How a Time-Limited Welfare Policy Affects Families at Risk of Reaching Time Limits. *Developmental Psychology*, 45(2), 383-400. DOI: 10.1037/a0014960

Moulton, Shawn, with Laura R. Peck & Stephen H. Bell. (2014). *Social Policy Impact Pathfinder (SPI-Path) Analytic Suite: SPI-Path|Individual User Guide*. Bethesda, MD: Abt Associates Inc.

Patton, Michael Quinn. (2010). Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use. New York, NY: Guilford Press.

Peck, Laura R. (2003). Subgroup Analysis in Social Experiments: Measuring Program Impacts Based on Post Treatment Choice. *American Journal of Evaluation*, 24(2), 157-187. DOI: 10.1016/S1098-2140(03)00031-6

Peck, Laura R. (2005). Using Cluster Analysis in Program Evaluation. *Evaluation Review*, 29(2), 178-196. DOI: 10.1177/01933841X04266335

Peck, Laura R. (2007). What are the Effects of Welfare Sanction Policies? Or, Using Propensity Scores as a Subgroup Indicator to Learn More from Social Experiments. *American Journal of Evaluation,* 28(3), 256-274. DOI: 10.1177/1098214007304129

Peck, Laura R. (2013). On Analysis of Symmetrically-Predicted Endogenous Subgroups: Part One of a Method Note in Three Parts. *American Journal of Evaluation*, 34(2): 225-236. DOI: 10.1177/1098214013481666

Peck, Laura R. and Stephen H. Bell. (2014). The Role of Program Quality in Determining Head Start's Impact on Child Development. OPRE Report #2014-10, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Peck, Laura R., & Shawn Moulton. (2013). On The Use of Instrumental Variables and Symmetric-Prediction for Estimating Impacts of Mediators. Paper presented at the Welfare Research and Evaluation Conference, Washington, DC, May 30.

Raudenbush, Stephen W. & Sally Sadoff. (2008). Statistical Inference When Classroom Quality is Measured With Error. *Journal of Research on Educational Effectiveness*. 1(2): 138-154. DOI: 10.1080/19345740801982104

Schochet, Peter Z., & John Burghardt,. (2007). Using Propensity Scoring to Estimate Program-Related Subgroup Impacts in Experimental Program Evaluations. *Evaluation Review*, 31(2), 95-120.

Unlu, Fatih, Ryoko Yamaguchi, Larry Bernstein, Julie Edmunds. (2011). Estimating Impacts on Program-Related Subgroups in North Carolina's Early College High School Study. Cambridge, MA: Abt Associates Inc. Unpublished manuscript.

Unlu, Fatih, Laurie Bozzi, Carolyn Layzer, Arthur Smith, Cristopher Price, & R. Hurtig. (2013). Linking Implementation Fidelity to Outcomes in an RCT. Cambridge, MA: Abt Associates Inc. Unpublished manuscript.

Yoshikawa, H., E.A. Rosman, & Joann HsuehJ. (2001). Variation in Teenage Mothers' Experiences of Child Care and Other Components of Welfare Reform: Selection Processes and Developmental Consequences. *Child Development*, 72, 299–317.

Zanutto, Elaine, Bo Lu, & Robert Hornik. (2005). Using Propensity Score Subclassification for Multiple Treatment Doses to Evaluate a National Antidrug Media Campaign. *Journal Of Educational And Behavioral Statistics*,

30(1), 59-73. DOI: 10.3102/10769986030001059

Zhai, Fuhua, C. Cybele Raver, & Stephanie M. Jones. (2012). Academic Performance of Subsequent Schools and Impacts of Early Interventions: Evidence from a Randomized Controlled Trial in Head Start Settings. *Children and Youth Services Review*, 34(5), 946–954. DOI: i:10.1016/j.childyouth.2012.01.026

Zhai, Fuhua, C. Cybele Raver, & Christine Li-Grining. (2011). Classroom-based Interventions and Teachers' Perceived Job Stressors and Confidence: Evidence from a Randomized Trial in Head Start Settings. *Early Childhood Research Quarterly*, 26(4), 442-452. DOI: 10.1016/j.ecresq.2011.03.003

Zhai, Fuhua, C. Cybele Raver, Stephanie M. Jones, Christine P. Li-Grining, Emily Pressler, & Qin Gao. (2010). Dosage Effects on School Readiness: Evidence from a Randomized Classroom-Based Intervention. *Social Service Review,* 84(4), 615-655. DOI: 10.1086/657988