Article

American Journal of Evaluation
1-18
© The Author(s) 2015
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1098214015594419
aje.sagepub.com
SAGE

Principal Stratification: A Tool for Understanding Variation in Program Effects Across Endogenous Subgroups

Lindsay C. Page[1], Avi Feller[2], Todd Grindal[4], Luke Miratrix[3], and Marie-Andree Somers[5]

Abstract

Increasingly, researchers are interested in questions regarding treatment-effect variation across partially or fully latent subgroups defined not by pretreatment characteristics but by post-randomization actions. One promising approach to address such questions is principal stratification. Under this framework, a researcher defines endogenous subgroups, or principal strata, based on post-randomization behaviors under both the observed and the counterfactual experimental conditions. These principal strata give structure to such research questions and provide a framework for determining estimation strategies to obtain desired effect estimates. This article provides a nontechnical primer to principal stratification. We review selected applications to highlight the breadth of substantive questions and methodological issues that this method can inform. We then discuss its relationship to instrumental variables analysis to address binary noncompliance in an experimental context and highlight how the framework can be generalized to handle more complex posttreatment patterns. We emphasize the counterfactual logic fundamental to principal stratification and the key assumptions that render analytic challenges more tractable. We briefly discuss technical aspects of estimation procedures, providing a short guide for interested readers.

Keywords

causal models, quantitative methods, RCT, impact evaluation

Introduction

A truism of social science research is that everything varies. While we cannot expect to understand all aspects of impact variation, we can often learn about variation in effects across particular

[1] University of Pittsburgh, Pittsburgh, PA, USA
[2] University of California, Berkeley, CA, USA
[3] Harvard University, Cambridge, MA, USA
[4] Abt Associates, Cambridge, MA, USA
[5] MDRC, Los Angeles, CA, USA

Corresponding Author:
Lindsay C. Page, University of Pittsburgh, 5918 WWPH, 230 S. Bouquet Street, Pittsburgh, PA 15260, USA.
Email: lpage@pitt.edu

subpopulations. For example, we might estimate how the effect of an intervention varies across certain observable individual characteristics—such as gender, race/ethnicity, or socioeconomic status. In the context of a randomized controlled trial (RCT), estimating these subgroup treatment effects is typically a straightforward analytic exercise, often completed by estimating a treatment-control difference within the subsample defined by the subgroup characteristic of interest or by including interaction terms between treatment and subgroup indicators in a linear model.[1] These methods give an estimated treatment effect for each subgroup, and testing procedures such as likelihood ratio tests can determine whether the subgroup effects are significantly different. In essence, these approaches view the subgroups of interest as mini-experiments and return estimates of treatment impact for each such experiment.

Increasingly, however, researchers are interested in treatment-effect variation across partially observed—or even fully latent—subgroups, defined not by standard pretreatment characteristics, such as gender or race/ethnicity, but instead by post-randomization behaviors, actions, or decisions. For example:

1. What is the impact of the offer of a private school voucher among those children who would enroll in private school only when given a voucher, but who would otherwise enroll in public school?
2. What is the impact of a multifaceted dropout prevention program among those students who would fail to complete high school regardless of whether they have the opportunity to participate in the program? In contrast, does the program only improve outcomes for students who graduated from high school as a result of the opportunity to participate in the program?
3. What is the impact of participating in an early childhood program among those children who would participate in the program when offered the opportunity to enroll in the program but who would have received home-based care absent that opportunity? And how does this compare to the impact for those children who would have otherwise enrolled in some other center-based care?

As compared to standard subgroup analysis, in these examples we are not able to observe individual membership in the subgroups of interest, thus preventing classic subgroup approaches. Two key features are worth highlighting about the subgroups of interest identified in these research questions. First, the subgroups are defined by specific post-randomization behaviors. Second, to classify study participants into these subgroups of interest, we would need to be able to observe individual-level behavior under both the observed and the unobserved (i.e., counterfactual) experimental conditions. In the first question mentioned previously, we are interested in those children who would enroll in public school under assignment to the control condition but who would enroll in private school upon receipt of the randomly assigned voucher. In the second, we are interested in those students who would fail to graduate from high school regardless of the experimental condition to which they are assigned. In the third, we are interested in those children who would take up the early childhood program if offered, grouped by the care setting that these children would experience if assigned to the control condition. All these types of questions allow us to further unpack overall intent-to-treat (ITT) effects to understand for whom and under what circumstances a given intervention impacts outcomes for children and families. Answers could help to determine whether observed ITT effects are driven by one particular latent subgroup, and if so, how the intervention might be restructured to better serve the full range of intended participants.

In reality, of course, we can only observe a study participant's post-randomization behavior under assignment to treatment or control—but not both. One analytic framework for handling such research questions is *principal stratification* (Frangakis & Rubin, 2002). The idea behind principal stratification is to first define endogenous (or program-related) subgroups, referred to as *principal strata*, based on sample members' post-randomization choices, actions, or experiences under both the observed and the counterfactual experimental conditions and then to leverage statistical methods to estimate impacts for these subgroups, even though they are not fully observed.

Like many statistical concepts, principal stratification has several, independent origins. In the program evaluation literature, the idea appears to originate in the analysis of the Job Training Partnership Act (Bloom et al., 1997), with further discussion in Orr (1999), and eventually as *endogenous* or *program-related subgroups* (Peck, 2003; see Peck, 2013, for a discussion). In statistics, the term principal stratification arose from complications in medical trials and was eventually formalized by Frangakis and Rubin (2002). A particular contribution of this framework is that it provides clarity in differentiating the process of defining treatment effects of interest from the analytic strategies used to estimate those quantities, as we illustrate further below. Consequently, researchers can articulate treatment effects using natural language to define and describe specific groups of interest while also clarifying the assumptions necessary to estimate treatment effects for those groups. Importantly, in this framework, stratum membership is considered a pretreatment characteristic of each study participant—just like age or socioeconomic status—because the participants' set of potential responses to the different treatments is also considered to be a fixed, pretreatment characteristic. Thus, stratum-specific treatment effects, or *principal causal effects*, are subgroup treatment effects, just like treatment effects for other subgroups, such as boys and girls.

This article is intended as a nontechnical primer on principal stratification. In the second section, we highlight several different applications of the principal stratification framework to illustrate the breadth of substantive questions and methodological issues to which it can be applied. In the third section, we focus on key assumptions that underlie our identification and estimation of stratum-specific treatment effects. In the fourth section, we present two different approaches to estimating key principal causal effects in the context of a randomized trial that suffers from simple (i.e., binary) noncompliance, and highlight how certain estimation procedures can be extended to handle more complex applications of the principal stratification framework. We close by highlighting the benefits and limitations of the principal stratification framework as well as the current state of the field regarding analytic tools that would be of interest and value to the applied quantitative researcher. Throughout, we emphasize the counterfactual logic fundamental to this type of analysis as well as the assumptions that undergird this approach and render analytic challenges more tractable. Although we touch on estimation procedures, they are not the primary focus.

Selected Applications of Principal Stratification in the Social Sciences

Researchers have utilized the principal stratification framework in the context of RCTs to define treatment effects of interest, while rigorously handling analytic challenges and addressing substantive questions related to many concerns. Some examples, discussed below, are simple and complex patterns of noncompliance to treatment assignment (e.g., Barnard, Frangakis, Hill, & Rubin, 2003); naturally occurring variation in the counterfactual, nontreatment condition (Feller, Grindal, Miratrix, & Page, 2014); missing or unobserved intermediate outcomes (e.g., Frumento, Mealli, Pacini, & Rubin, 2012); and the identification of shorter term "surrogate" outcomes of longer term outcomes of interest (VanderWeele, 2011). We highlight recent applications in each of these domains drawn from research in education and workforce development to provide a sense for the breadth of questions to which researchers have fruitfully applied the principal stratification framework.

Simple and Complex Patterns of Noncompliance With Assignment to Treatment

In experimental studies in the social sciences, study participants often do not comply with the randomized assignment to an active intervention. Those assigned to the treatment condition may fail to take up the treatment, and those assigned to the control condition may nevertheless gain access to the treatment. Barnard, Frangakis, Hill, and Rubin (2003), for example, encountered this challenge in their analysis of the New York City (NYC) School Choice Scholarship Program. Through this intervention, eligible students were randomly selected to receive scholarship vouchers to help cover the cost of private (primarily parochial) school enrollment in NYC. In this context, students could react to random assignment in different ways. For those randomly offered the voucher (the treatment group), students could either use the voucher to enroll in a private school or decline the voucher and enroll in public school. For those not offered a voucher (the control group), students could either enroll in private school regardless of being denied the financial support or enroll in public school.

Figure 1 shows these possible responses to random assignment, as originally articulated by Angrist, Imbens, and Rubin (1996). In this figure, the row headings correspond to how students would respond under assignment to the voucher (treatment) condition—either they enroll in private school or not. The column headings similarly indicate students' response under the no-voucher (control) condition. The four possible pairs of potential responses under assignment to the treatment and control conditions categorize individuals as falling into one of the four response "profiles," our principal strata of interest in this context, as shown in Figure 1.

A key goal in this intervention was to understand the impact of private schooling on students' educational outcomes. Analytically, therefore, there is substantive interest in the impact of treatment within the subgroup of students who are "compliers," those who only enroll in private school when offered the voucher. For reference, the other groups that are definitionally possible are "always takers," students who would enroll in private school regardless of assignment; "never takers," students who would never enroll in private school regardless of their assignment; and "defiers," students who would enroll in public school if offered the voucher but who would enroll in private school without the voucher offer.[2] In this case, Barnard and colleagues (2003) find that the voucher offers improved mathematics outcomes among those compliers who attended schools performing below the citywide median prior to the study.

Because imperfect compliance is typical in social science experiments, this goal of estimating the treatment effect for the subset of individuals who participate in the experiment as intended is fairly common. In addition to understanding impacts among compliers, this framing also encourages the policy maker to consider why certain individuals did not respond according to the intention of the experimental design. Indeed, viewing simple noncompliance in this way is now standard practice (e.g., Abdulkadiroglu, Angrist, Dynarski, Kane, & Pathak, 2011; Bloom, 1984; Howell, Wolf, Campbell, & Peterson, 2002).

However, more complex noncompliance patterns are also possible. For example, if an experiment is conducted over an extended time period, it is possible that individuals' compliance with assigned treatment may not be simply an either/or phenomenon. Rather, compliance to the assigned treatment may vary over time. We can still use a principal stratification framework to handle these more complex patterns of noncompliance, as was done by Jin and Rubin (2009) to extend analyses of the NYC voucher experiment to consider the study's multiyear timeframe.

Here the authors observed that even among voucher recipients who took up the voucher offer in the first year, some reverted to attending public school over the subsequent years of the study. Such patterns created a need to define principal strata for varying degrees of compliance over time, known as partial noncompliance. In this context, strata were defined by the various patterns of private school enrollment students could have exhibited over a 3-year period under both experimental conditions. Structuring the analysis in this way allowed the authors to isolate effects for children who would take up the voucher and attend private school for different durations and to explore a variety of hypotheses, such as whether the voucher offer has a negative impact on those who would only partially comply with assignment due to an "adjustment hardship" that they might experience from multiple changes in school environment.

Figure 1. Cross-classification of treatment take-up behavior for defining principal strata in study of vouchers to support private school enrollment.

		Random assignment to no-voucher condition (control)	
		Private school enrollment	Public school enrollment
Random assignment to voucher condition (treatment)	Private school enrollment	1: Always (private school) takers	2: Compliers
	Public school enrollment	3: Defiers	4: Never (private school) takers

Variation in Counterfactual Condition

Another use of principal stratification recently forwarded by Feller, Grindal, Miratrix, and Page (2014) is to investigate naturally occurring variation in counterfactual conditions. Using data from the Head Start Impact Study (HSIS), the authors observed that children not offered the opportunity to enroll in Head Start experienced different forms of child care such as alternate, non-Head Start child care centers, or care at home provided by a parent or other relative. This variation motivated the question of whether there was variation in the impact of Head Start participation according to the care setting that children would have experienced absent the Head Start offer.

Here the principal strata of interest are defined by the care setting children would have experienced under assignment to treatment and under assignment to control, as illustrated in Figure 2. These principal strata in Figure 2 generalize the noncompliance represented in Figure 1. In this application, the key subgroups are the two types of compliers: those compliers who would otherwise experience center-based care (Cell 2 in Figure 2) and those compliers who would otherwise experience home-based care (Cell 3). Importantly, Feller and colleagues find that the main treatment effects of the Head Start intervention were largely realized by those children who were induced into Head Start and out of a home-based care setting as a result of the randomized offer to enroll, whereas the impact of enrolling in Head Start was essentially zero for those children who would otherwise have been in another non-Head Start center-based setting. Such results have implications regarding determinations of program effectiveness and the targeting of program expansion.

Handling Unobserved or Undefined Outcomes

Another circumstance in which principal stratification can be useful occurs when an outcome of interest is only observable for the subset of individuals who experienced another, related outcome. For example, consider interventions that focus on workforce development, such as Job Corps (e.g., Frumento et al., 2012; Zhang, Rubin, & Mealli, 2008). In assessing the impact of such programs, we may have a particular interest in how assignment to the program affects wages. The complication is that wages are only observable and well defined for individuals who are employed. One naive approach would simply be to compare wages among all individuals observed to be employed post-treatment. However, an obvious problem with this approach is that the intervention might induce into the labor market those who earn particularly low wages when employed. If this is the case, we might conclude incorrectly that the intervention has a negative impact on wages. Another option is to assign a value of zero for the wages of those who are not working. This decision, however, would lead to the estimation of the program's impact on a combination of employment and wages conditional on employment.

Figure 2. Cross-classification of child care setting experienced for defining principal strata in study of variation in the impact of Head Start according to type of care otherwise experienced.

		Random assignment to no Head Start offer (control)		
		Head Start	Center Care	Home Care
Random assignment to Head Start offer (treatment)	Head Start	1: Always Head Start	2: Center Compliers	3: Home Complier
	Center Care	(A)	4: Always Center Care	(B)
	Home care	(C)	(D)	5: Always Home Care

Principal stratification offers a useful solution to this common analytic problem. Zhang, Rubin, and Mealli (2008) use a principal stratification framework and define strata according to employment status under assignment to treatment and control, as illustrated in Figure 3. Here, individuals can belong to one of the four possible groups: those who would always be employed regardless of treatment, those who would never be employed regardless of treatment, those who would be employed only if assigned to treatment, and those who would be employed only if assigned to the control condition. This framework allows the authors to estimate the effects that are of particular interest in understanding the impacts of the program. First, by estimating the share of participants in each stratum, the authors estimate the impact of the intervention on employment. They do so by comparing the share of individuals in two strata: those who would be employed only if treated (i.e., individuals for whom the program had a positive impact on employment) and those who would be employed only if not treated (i.e., individuals for whom the program had a negative impact on employment).[3] The difference between the share in the second and the share in the first is the impact on employment. Second, for those who would be employed under either experimental condition, they can examine the impact of treatment assignment on wages. In related work, Lee (2009), for example, finds that Job Corps does lead to an increase in wages for those who would be employed under either experimental conditions and, based on these results, concludes that the program impacts labor market outcomes both by increasing human capital as measured by increased wages among this subset of study participants and by increasing rates of employment among those who would not be employed absent the opportunity to participate in Job Corps.

This type of application of the principal stratification framework, involving unobserved or undefined outcomes (e.g., wages for those who are unemployed), is sometimes referred to as the "truncation by death" problem (see also McConnell, Stuart, & Devaney, 2008, for an excellent review; Rubin, 2006; Zhang & Rubin, 2003). This terminology arose from the original application of principal stratification to quality of life studies where the analytic challenge is that quality of life is not well defined for those study subjects who have passed away during the course of the intervention or follow-up period.

Surrogate Outcomes and Mediation

A third application of principal stratification involves surrogate outcomes and mediation. Surrogate outcomes become important when it is unduly expensive or infeasible to track longer run outcomes of interest. Questions of mediation are important for efforts to understand causal pathways by which interventions operate. One recent application in this domain is Page's (2012) study of career academy high schools. MDRC's experimental evaluation of this high school model found that the randomized offer to enroll in a career academy had no effects on traditional educational outcomes, such

Figure 3. Cross-classification of employment behavior for defining principal strata when the outcome of interest is observed conditional on employment.

		Random assignment to no job training program (control)	
		Employed	Not Employed
Random assignment to job training program (treatment)	Employed	1: Always employed	2: Employed if treated
	Not employed	3: Employed if not treated	4: Never employed

as high school performance, high school completion, or college attainment. Yet, several years after high school, students randomized to a career academy had substantially higher earnings than their control group counterparts. This set of results presented a puzzle, given that we often look to these traditional educational milestones as important pathways to subsequent labor market success. Page (2012) explored the hypothesis that exposure to the world of work through opportunities such as internships and job shadowing provided by the program contributed to these positive effects. To explore this hypothesis, a key analytic step is to stratify students according to the extent of the change in labor market exposure they would have experienced if given the opportunity to participate in the career academy. The strata of interest for this analysis, therefore, were defined by the level of world-of-work exposure students would have received under the treatment and control conditions. Based on student-reported information on participation in labor-market exposure activities, Page categorized students into low, moderate, and high levels of exposure in both the treatment and the control conditions.[4] The strata of interest in this application are illustrated in Figure 4. By estimating treatment effects within each stratum, Page finds that treatment effects on subsequent earnings are largest among those students who also experienced the largest change in exposure to the world of work as a result of the career academy offer.[5] This finding is relevant because it is consistent with hypotheses regarding the key programmatic components of career academies that lead to subsequent labor market success for students. Nevertheless, concluding that a surrogate outcome is a mediator involves making further assumptions regarding mechanism. Currently, there is some debate as to how to do this within a principal stratification framework.[6]

Applying Assumptions to Incorporate Knowledge and Support Estimation

A principal stratification-based analysis proceeds in three stages. The first stage is to identify the principal strata as we described earlier. In general, this process involves thinking about the substantive questions in a relatively high-level manner. The next stage is to formalize these strata and also encode further substantive information that can ease later estimation. We discuss these first two stages in this section. The final stage is estimation, discussed in the fourth section.

First, for strata to be well defined, we need to make some assumptions as to how the randomized trial works. Principal stratification is based on the potential outcomes framework (Frangakis & Rubin, 2002; Neyman, 1923; Rubin, 2005), which effectively considers each study individual to have individual potential outcomes and associated potential treatment effect. The potential outcomes of person i, $Y_i(1)$ and $Y_i(0)$ are the outcomes we potentially would see if we treated or did not treat the person, respectively. Given these values, a key quantity of interest, the overall average treatment effect, is the average of these individual treatment effects. This framework has great power in that

Figure 4. Cross-classification of world-of-work exposure for defining principal strata in study of impact of the career academy high school model on post-high school earnings.

		Random assignment to no career academy offer (control)		
		Low labor market exposure	Moderate labor market exposure	High labor market exposure
Random assignment to Career academy offer(treatment)	Low labor market exposure	Always low	(A)	(B)
	Moderate labor market exposure	Low to moderate	Always moderate	(C)
	High labor market exposure	Low to high	Moderate to high	Always high

the causal impact is well defined and does not rely on any distributional or sampling assumptions: the population consists of the individuals in the experiment and no more. The challenge is that while all the potential outcomes are defined, for each person, only one potential outcome is actually observed, according to the randomized treatment assignment.

To progress from these basic definitions, we first make two generally accepted assumptions. The first is that the treatment was in fact randomized, which we express as an independence assumption between whether an individual gets treated and what the person's potential outcomes are. The second, commonly called the stable unit treatment value assumption (SUTVA) (Rubin, 1978, 1980, 1990), effectively says that treating one person does not impact another (i.e., treating one student has no spillover effect that helps some other student). For those not familiar with the potential outcomes framework, we suggest Angrist et al.' study (1996), which discusses this in the case of noncompliance. For further discussion, see Imbens and Rubin (2015) or the first third of Rosenbaum (2010).

The above-mentioned assumptions give us well-defined strata and well-defined treatment effects within each stratum. The reasoning is as follows: given SUTVA, how an individual would respond to either experimental condition is essentially a characteristic of the individual. This, in turn, implies that each individual's stratum of membership is in essence a characteristic of the individual. Given that treatment assignment is randomized, it is unrelated to stratum membership. Therefore, within-stratum treatment effects are causal effects, just as other subgroup treatment effects would be. With strata and stratum-level treatment effects defined, we move to the next stage where we add substantive assumptions to place constraints on the strata. These restrictions are important as, without which, it is often difficult to make progress toward estimating the key quantities of interest. More importantly, assumptions allow us to formally incorporate our substantive knowledge of the problem into our analyses. This is a particular benefit of the principal stratification framework. We briefly discuss the two most common assumptions in the context of standard noncompliance, monotonicity, and exclusion restrictions, before discussing the other settings for principal stratification mentioned earlier.[7]

Monotonicity and Exclusion Restrictions in the Context of Binary Noncompliance

The two common assumptions in the context of noncompliance in randomized experiments are *monotonicity* and the *exclusion restriction* (Angrist, Imbens, & Rubin, 1996). Monotonicity relates to how people actually respond to the randomized offer to take up a given intervention. Monotonicity is the assumption that an individual would be at least as likely to take up the treatment under assignment to the treatment condition as under assignment to the control condition. Returning to Figure 1, which presents the principal stratification set up for binary noncompliance in the context of the NYC school voucher experiment, the monotonicity assumption rules out the possibility of any "defiers," as children in this stratum would have enrolled in private school without the voucher offer but not with the voucher offer.

The second major assumption is the *exclusion restriction*, which relates to how assignment to the treatment condition will impact those individuals who would not actually take up the treatment, if offered. Again, in the context of the example represented in Figure 1, these are the children whose school type would be unchanged by the randomized voucher offer. The always takers would always enroll in private school, with or without the voucher, whereas the never takers would always enroll in public school. Given that randomization does not change the type of schooling for these subgroups of children, we make the assumption that randomization similarly does not impact subsequent outcomes for these children. Equivalently, we assume that, within these strata, the effect of treatment is zero.

Importantly, these assumptions are inherently substantive rather than statistical. Therefore, it is critical to reason through whether such assumptions are plausible in the context of each substantive application. In the context of the voucher study, monotonicity seems entirely plausible, as it is hard to imagine a student or family being more likely to take up private schooling without the financial assistance that the voucher provides. The exclusion restrictions for always takers and never takers, however, merit further consideration. Never takers are those children who would attend public school regardless of the voucher offer. Because school setting and, quite likely, actual school building were unchanged for these children, it is reasonable to assume that the treatment offer did not impact subsequent school-related outcomes. For always takers, however, it is conceivable that the voucher offer could have led families to select different private schools for their children. If the voucher allowed some families to opt into higher quality private school settings, the treatment effect among always takers could plausibly be nonzero. Given this, research must rigorously consider and defend the application of the exclusion restriction for always takers or not apply this assumption and instead estimate a treatment effect for this subgroup. These assumptions can in some cases be tested, as we discuss in the estimation section below.

Assumptions Applied to Other Principal Stratification Examples

To further highlight the use of these assumptions, we briefly discuss their application in the context of the other examples noted above. First, consider the examination of the impact of Head Start according to alternative care settings, as represented in Figure 2 (Feller et al., 2014). The monotonicity assumption in this context implies that children would be at least as likely to participate in Head Start if offered a slot than if not offered a slot. By applying this assumption, we rule out the possibility of children belonging to the strata labeled as (A) and (C), as these are children who would participate in Head Start under assignment to the control condition but who would take up other center-based care or home-based care, respectively, if offered the opportunity to enroll in Head Start. Feller and colleagues (2014) also rule out the possibility of children belonging to the strata labeled (B) and (D). Analogous to a monotonicity assumption, the authors make the assumption that the Head Start offer would not induce families to switch from, say, center-based care to a home-based care setting or vice versa.

Further, the authors apply an exclusion restriction to the cells on the main diagonal of Figure 2, making the assumption that the effect of the treatment offer is zero among those children whose child care sector would be unchanged by the randomized offer to participate in Head Start. As in the application above, it is useful to consider scenarios under which this assumption would be violated. Consider, for example, the children belonging to the always Head Start stratum. It is possible that without the Head Start offer, these children could have participated in a low-quality Head Start center but that the randomized offer corresponded to the opportunity to enroll in a high-quality Head Start center. Under such circumstances, it is plausible that these children would experience a positive effect of treatment on subsequent outcomes. In the HSIS, however, these control group children often enrolled in the very same Head Start centers to which their access was supposedly denied. This understanding of the roll out of the HSIS, therefore, justifies this application of the exclusion restriction.

In the wage example, neither monotonicity nor exclusion restriction assumptions seem plausible. Regarding monotonicity, participation in the job training program may increase expectations for wage and job quality and may lead individuals to engage in a longer job search (see Note 3). As a result, participating in the program might *decrease* the probability of having a job. Therefore, in this context, it is possible for all four strata to exist. Regarding the exclusion restriction (i.e., the assumption of zero impact within certain strata), there is only one stratum—the always employed—within which a treatment effect on wages will be estimated, and the effect of the treatment on wages is undefined in all other cells.

Finally, we consider Page's (2012) application of principal stratification to the context of career academy high schools. In this context, Page invokes a monotonicity assumption, arguing that participating in labor market exposure activities would be at least as high when offered the opportunity to enroll in a career academy. This assumption rules out the possibility of students belonging to the strata (A), (B), or (C) in Figure 4. She does not, however, apply an exclusion restriction, given that the intervention could have influenced students' outcomes through mechanisms other than labor market exposure. Therefore, she allows treatment effects to be nonzero for students whose level of labor market exposure would be unchanged by the opportunity to participate in a career academy but finds impact estimates within these strata to be small compared to those strata within which labor market exposure increased with the career academy offer.

Strategies for Estimating Stratum-Specific Treatment Effects

As the above-mentioned discussion indicates, principal strata are defined and substantive assumptions to restrict the characteristics of these strata are applied without reference or prior to estimation. In our view, this separation is a beneficial aspect of utilizing the principal stratification framework: the first articulate the groups and quantities of interest and the second identify the substantive knowledge (e.g., assumptions) that could be useful to further restrict or inform these quantities, and only then assess whether and how estimation is possible. In short, the process of estimation is separate and distinct from the process of articulating those quantities to be estimated.

Depending on the constraints imposed on the problem, there are different general strategies one might take for estimation. These strategies can be divided into moment-based methods (nonparametric) and model-based methods. In the first, we rely purely on our original assumptions to express target quantities (such as stratum-specific average treatment effects) as functions of directly observable characteristics of the data. When this is not possible, moment-based approaches can be weakened to instead attempt to bound target quantities. The second approach is to directly model part or all of the data. This typically requires making distributional assumptions on the outcomes within the principal strata and also utilizing baseline covariates to inform prediction of individuals' stratum membership.

We refer to the first approach as "moment-based," because we can estimate relevant stratum-specific treatment effects based on estimable "moments" or quantities such as means and proportions. The second approach is "model-based," because our estimation procedure involves estimating parameters from an articulated model (Imbens & Rubin, 1997). The key difference between these two approaches is that the moment-based approach only uses sample-level information while the model-based approach uses individual-level information, as we discuss below. In the context of noncompliance, the first strategy may be more familiar to the applied quantitative researcher, while the tools and procedures associated with the second strategy are often necessary for estimating principal causal effects in more complex applications of principal stratification. For a textbook discussion comparing these two approaches, see Imbens and Rubin (2015). Also see the introduction of Stuart, Perry, Le, and Ialongo (2008).

These two general approaches have different strengths and weaknesses. Moment-based, nonparametric methods rely on weaker assumptions and allow for nonstandard and unknown distributional forms of the outcomes under consideration. However, this lack of structure makes disentangling many strata more difficult. Model-based methods can, at least in principle, separate out multiple strata to allow for insight into complex patterns of treatment impact. Importantly, however, the fundamental analytic exercise is one of separating mixtures—a notoriously difficult problem—and doing this can be somewhat of a delicate business, with treatment effect estimates potentially being sensitive to modeling choices.

To illustrate the conceptual underpinnings, we walk through both approaches for the now well-understood case of simple noncompliance (such as the voucher study represented in Figure 1). Here, using the moment-based method together with the monotonicity and exclusion restriction assumptions, we can use instrumental variables (IVs) estimation to estimate the treatment effect among compliers (Angrist et al., 1996). We then provide intuition for applying a model-based approach to this analytic problem.

Moment-Based IVs

For moment methods, we are usually concerned with differences in means for different groups. In the context of noncompliance, we typically denote the overall difference in treatment and control group means for a given outcome (or the overall impact of randomization) as the ITT effect, and the difference in treatment and control group means specific to the compliers as ITT_c. This latter quantity is also referred to as the complier average causal effect or local average treatment effect.[8]

In general, we can rewrite an overall treatment effect as a weighted average of subgroup-specific effects. For example, if we were examining the impact of an intervention on high school students, we can decompose the overall effect into the weighted average of the effect on students in each grade, 9 through 12:

$$\mathrm{ITT} = \pi_9 \mathrm{ITT}_9 + \pi_{10} \mathrm{ITT}_{10} + \pi_{11} \mathrm{ITT}_{11} + \pi_{12} \mathrm{ITT}_{12},$$

where π_g denotes the proportion of all students in Grade g, and ITT_g is the impact of randomization for students in Grade g.

In the context of noncompliance, we instead decompose the overall treatment effect into the effect on each of the four principal strata represented in Figure 1:

$$\mathrm{ITT} = \pi_c \mathrm{ITT}_c + \pi_a \mathrm{ITT}_a + \pi_n \mathrm{ITT}_n + \pi_d \mathrm{ITT}_d, \tag{1}$$

where the c, a, n, and d subscripts refer to compliers, always takers, never takers, and defiers, respectively. As shown by Angrist et al. (1996), with the monotonicity and exclusion restrictions, we are able to derive the ITT_c, the principal causal effect for compliers, even though we do not know which individual is in which group. First, recall that with the application of the monotonicity assumption, we rule out the existence of defiers. In Equation 1, this assumption corresponds to $\pi_d = 0$ Second, with the application of exclusion restrictions, we assume a treatment effect of zero for always takers and never takers. In Equation 1, this pair of restrictions corresponds to $\mathrm{ITT}_a = 0$ and $\mathrm{ITT}_n = 0$. With these additional restrictions, Equation 1 reduces to the following:

$$\mathrm{ITT} = \pi_c \mathrm{ITT}_c + \pi_a \underbrace{\mathrm{ITT}_a}_{0} + \pi_n \underbrace{\mathrm{ITT}_n}_{0} + \underbrace{\pi_d}_{0} \mathrm{ITT}_d = \pi_c \mathrm{ITT}_c. \tag{2}$$

Next, we estimate both ITT and π_c directly from the data. First, we estimate the overall ITT as the difference in mean outcomes between the treatment and the control groups. Second, we estimate π_c as the observed share of individuals in the treatment group who take up the treatment (which includes compliers and always takers, $\hat{\pi}_c$ and $\hat{\pi}_a$) minus the share of individuals in the control group who take up treatment (always takers, $\hat{\pi}_a$). We then scale the overall ITT estimate by the estimated share of compliers to obtain our final estimate. This is called the ratio estimator or the IV estimator for the effect of treatment on the subset of compliers.

Model-Based IVs

An alternative approach is to focus on each individual's stratum membership. Of course, we cannot observe this directly, since each individual is assigned to either treatment or control. Still, we do observe partial information. For example, in the voucher study, if a child is randomized to receive the voucher and subsequently enrolls in private school, we know that the child is either a complier or an always taker. At the same time, if a child receives the voucher but still enrolls in public school, then, by assumption, we know that we have a never taker. These possible relationships are shown in Table 1.

We refer interested readers to Imbens and Rubin (2015) for a detailed discussion of model-based IV. To give some intuition, we provide a brief sketch of the Bayesian estimation method here, which is known as data augmentation. Data augmentation is based on the idea that if we knew each individual's stratum membership, estimating the effect for compliers would be easy—we would simply estimate the treatment effect for the subgroup of individuals known to be compliers. The basic idea, therefore, is to predict stratum membership and then proceed *as if* that membership was known. Since there is uncertainty in this prediction, we repeat this procedure many times, via Markov chain Monte Carlo.

Since this is model-based estimation, we need to impose a model on the data. In this example, we assume that the outcome distribution for each principal stratum follows a normal distribution with mean μ_s and variance σ_s^2, which we write as $y_i \mid S_i = s \sim N(\mu_s, \sigma_s^2)$ for stratum s. Due to the exclusion restrictions, we assume that, for always takers and never takers, these model parameters are the same under treatment and control, so we have four parameters for these two groups: (μ_a, σ_a^2) and (μ_n, σ_n^2). For the compliers, we also have four parameters to estimate, two each under treatment and control: $(\mu_{c0}, \sigma_{c0}^2)$ and $(\mu_{c1}, \sigma_{c1}^2)$. We continue with the monotonicity assumption that defiers do not exist. Once we have estimated all these parameters, we can directly estimate the ITT_c, which is the simple difference in means between compliers in the treatment group and compliers in the control group $(\mu_{C1} - \mu_{C0})$.

The basic estimation strategy has two steps: (1) assuming we know stratum membership, estimate model parameters and (2) assuming we know the model parameters, predict stratum membership. The first step is immediate—once we know each individual's stratum membership, we can simply estimate the mean and variance of the outcome within each stratum and also directly estimate the proportion of students within each stratum. The second step, however, is more complicated. For some students, such as those who are offered a voucher but still attend public school, compliance type is known: These students are never takers. However, as shown in Table 1, a student who is offered a voucher and attends private school could be either a complier or an always taker. In this instance, we essentially perform a weighted coin flip to determine the student's compliance type.

Table 1. Possible Principal Strata Under Monotonicity.

Treatment assigned	Treatment received	Principal stratum
0	0	Complier or never taker
0	1	Always taker
1	0	Never taker
1	1	Complier or always taker

The key question is how to determine the weights. The simplest option is to use the relative proportions of compliers and always takers in the overall sample:

$$P(\text{child } i \text{ is a complier}) = \frac{\hat{\pi}_c}{\hat{\pi}_c + \hat{\pi}_a}.$$

For example, if there are twice as many compliers as always takers, then the student is predicted to be a complier with the probability two thirds. However, what if compliers under treatment tend, on average, to have a higher value of the outcome of interest than always takers, that is, $\mu_{c1} > \mu_a$? Under this assumption, if student i has a relatively high outcome value (e.g., a standardized test score), then we believe that the student is more likely to be a complier than if the student had a relatively low test score. We can incorporate this information via Bayes' rule:

$$P(\text{student } i \text{ is a complier} \mid y_i, \overrightarrow{\theta}) = \frac{\hat{\pi}_c \times N(y_i;\ \mu_{c1}, \sigma^2_{c1})}{\hat{\pi}_c \times N(y_i;\ \mu_{c1}, \sigma^2_{c1}) + \hat{\pi}_a \times N(y_i;\ \mu_a, \sigma^2_a)},$$

where y_i is the outcome for student i, $\overrightarrow{\theta}$ is the vector of model parameters, and $N(y_i;\ \mu_s, \sigma^2_s)$ is the probability of observing outcome y_i from a Normal distribution with mean μ_s and variance σ^2_s. This ensures that a student's predicted membership also incorporates information from the outcome. Importantly, unless certain strong assumptions are met, failing to include this outcome information can lead to bias in the estimates of stratum-specific treatment effects.

These two steps, associated with "knowing" each individual's stratum membership and knowing the model parameters, are in essence the two key steps in our Bayesian estimation algorithm. We "start" the algorithm with initial guesses at the model parameters. Using these initial guesses, we predict the unknown stratum membership classification. Having assigned individuals to strata based on these predictions, we then estimate anew the model parameters. The algorithm iterates between these two steps while recording values of model parameters at each step.

Alternatively, instead of Bayesian approaches, one can implement data augmentation via the Expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). Both the EM approach and Bayesian approach rely on the idea of imputing the missing values of these latent (or partially observed) variables, but EM is a maximum likelihood-based approach. EM is often easier to implement (if the data are sufficiently informative about the model parameters of interest) using various stand-alone tools such as MPlus (Muthén & Muthén, 2010).

Adding Covariates

Up to this point, we have discussed the moment- and model-based estimation approaches without the inclusion of covariates. Of course, both can be extended to include covariates to improve the precision of estimates. For the first approach, this extension, known as two-stage least squares (TSLS), is ubiquitous in the social sciences. Although TSLS has some desirable robustness properties (e.g., Abadie, 2005), it can be difficult to interpret TSLS estimates in the presence of covariates—this only estimates the overall ITT$_c$ in special cases (Angrist & Pischke, 2008).

To add covariates to the model-based approach, we make the following changes to the modeling strategy. First, we allow π_s (the shares of participants in each strata) to depend on covariates. In this context, this means fitting a multinomial logistic regression model that predicts individual stratum membership based on covariates at each iteration. Second, we allow the parameters of the outcome distributions to depend on covariates. This similarly means modeling the means of the stratum-specific outcome distributions using a linear regression at each iteration, rather than simply taking sample means. For further discussion of the use of covariates in model-based IV, see, for example, Zigler and Belin (2012).

Sensitivity Checks

In general, if one can reasonably estimate the desired quantities with moment-based methods, we recommend using these approaches as a sensitivity check at the very least. While model-based methods are more powerful in that they can disentangle more complex effect structures and might have higher levels of precision, they are also more sensitive to modeling and functional form assumptions. Therefore, it is important to pressure test these models to see how much estimates change, given small perturbations and relaxations. For example, we advocate verifying that the point estimates from moment-based methods are more or less in line with the point estimates from model-based methods, up to measured uncertainty, if possible. One can also examine the sensitivity of results to model choices such as whether and which covariates are included. These sensitivity checks are distinct from those of relaxing underlying assumptions such as exclusion restrictions.

Nevertheless, assessing sensitivity to these identifying assumptions (e.g., exclusion restrictions and monotonicity) is also important and there are a variety of strategies to do so.[9] At root, the substantive justification and application of these types of assumptions lead to more tractable estimation. To check the validity of these assumptions, one can attempt to estimate the desired quantities with these assumptions relaxed in some form. If the overall pattern of results is consistent, then one has no cause for concern. As a final word, to paraphrase Grilli and Mealli (2007), results derived from model-based estimation must be viewed cautiously, as they are obtained through a process that invokes multiple assumptions. Therefore, an important first step in a principal stratification analysis is to estimate simple bounds on the key quantities of interest—how large or how small could the key treatment effects of interest be? It may be that the answer to this question would provide sufficient insight into the key research questions at hand.

Discussion

The principal stratification framework provides several benefits for considering important counterfactual questions such as those highlighted in the earlier examples. First, use of the framework requires great clarity in articulating the post-randomization experiences, decisions, or actions taken by experimental subjects under treatment and under control that are relevant to a given research question. This is useful in its own right. Second, having identified relevant strata, the framework additionally brings to the forefront the assumptions on which the estimation of treatment effects relies. For example, in the applications above, monotonicity assumptions ruled out the existence of certain strata and exclusion restrictions fixed treatment effects in certain strata to zero.

The principal stratification framework is beneficial for fostering clear thinking about defining the key quantities of interest. Yet, the process of estimation is comparatively less straightforward. That being said, some standard models are well understood, such as applications dealing with binary noncompliance and where the monotonicity and exclusion restriction assumptions are substantively defensible. Here moment-based IV is a classic solution, and model-based IV yields very similar estimates of the ITT_c (Imbens & Rubin, 1997). This equivalence is generally true in cases where there is a large sample size and a large proportion of compliers. In such cases, a reasonable question is why bother with the more complex model-based IV? One reason is that, if the proposed parametric model is approximately correct (here, most importantly, that the stratum outcome distributions are normally distributed), the model-based approach can lead to substantially more precise estimates (Imbens & Rubin, 1997). More importantly, however, the model-based approach provides a flexible framework within which to handle additional analytic complexity, such as missing data and the need to relax certain modeling assumptions.

For example, with a model-based approach, we can assess the sensitivity of the estimate of ITT_c to the exclusion restriction for never takers by allowing ITT_a to be nonzero. Recall the discussion above regarding the possibility that always takers who received a private school voucher would have the opportunity to attend a better private school. As would be expected, model stability is, in general, increased with the imposition of additional assumptions, but in many instances, this type of sensitivity check can be utilized to illustrate that the estimate of the ITT_c is not particularly sensitive to this exclusion restriction (Hirano, Imbens, Rubin, & Zhou, 2000). A final benefit of the model-based approach, and the one that makes it particularly important for some of the more complex applications of principal stratification discussed above, is that it is readily extended to settings in which the strata of interest are no longer defined by a binary variable (as in the Head Start and career academies examples above) and to settings in which monotonicity and exclusion restrictions are not valid assumption (as in the Job Corps example above).

There are potential drawbacks to this analytic strategy. First, model-based approaches can be sensitive to deviations from the model. Second, there is not a closed-form solution for the model-based IV estimate, which can make the estimation process seem opaque and difficult to explain to practitioners, as well as computationally difficult to execute. Finally, as we described previously, the model-based approach incorporates the outcome into the prediction of stratum membership. While sensible from a Bayesian perspective, this can be a hard sell. We point out, however, that sensitivity of results to modeling choices is not a problem specific to analytic strategies such as those discussed here. Nevertheless, for the many reasons that we highlight, sensitivity checks such as those described earlier are a key component of the analytic process.

Finally, it is important to note that the field has yet to progress to the point of providing applied researchers with the robust guidance and statistical software that would allow for broad utilization of model-based estimation strategies within the framework of principal stratification. In our own work, we have sought to make available tutorials and code that make it possible to replicate our own analyses (e.g., Feller et al., 2014) but recognize that more extensive adoption will require the development of analytic routines and user-friendly packages. Nevertheless, for readers interested in furthering their understanding of principal stratification, we recommend overviews provided by Imbens and Rubin (2015), Mealli and Mattei (2012), and Schochet, Puma, & Deke (2014) as a next set of references to investigate.

Authors' Note

We are grateful to Howard Bloom, Beth Boulay, Peng Ding, and Fabrizia Mealli for insightful feedback in writing this article. All opinions expressed in this article and any errors or omissions are our own.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article was supported by a grant from the Spencer Foundation entitled "Using Emerging

Methods with Existing Data from Multi-site Trials to Learn About and From Variation in Educational Program Effects.''

Notes

1. Throughout the article, we use the term ''treatment'' to refer generically to a program or intervention of interest.
2. As discussed further below, a common step when using the principal stratification framework is to apply content-specific knowledge to rule out the existence of certain strata, such as defiers in this instance.
3. At first blush, it may be hard to imagine a circumstance under which individuals would only be employed if they did not receive the intervention, since these are individuals for whom the impact of the program on employment would be negative. One possibility is that the training that individuals received through the program raised their expectations for the types of jobs to which they should have access. If a program raised such expectations disproportionate to the increase in skills for these individuals, the program may render some individuals less likely to be employed because they are less willing to take up the jobs for which they are able to be hired. If training leads individuals to seek jobs in fields where the search and hire period are longer or more involved, we may observe a short-run decline in employment as a result of the program. As we discuss later, a standard step in this type of analysis is to reason about which strata are likely to exist and to apply assumptions to reduce the number of strata. In this instance, it would be up to the analyst to gauge whether this stratum containing individuals who would be employed only without the job training program could possibly exist.
4. This highlights a limitation of the principal stratification framework that the variables utilized to define strata are typically categorical rather than continuous. See Jin and Rubin (2008) for one example of a principal stratification analysis utilizing a continuous measure to define strata.
5. Importantly, findings from this analysis are consistent with the hypothesized mechanism but are not able to provide conclusive evidence without further assumptions.
6. For a general discussion of mediation, see VanderWeele (2008) and Pearl (2011). A recent literature has sought to combine mediation with principal stratification (see, e.g., Jo 2008; Elliott, Raghunathan, & Li 2010; Imai, Jo, & Stuart 2011; Page 2012). Also see Rubin (2004) for a broader discussion of how principal stratification can help clarify investigation and interpretation of mediated (direct and indirect) effects.
7. While we discuss the most common assumptions here, there is a large literature on a broader set of possible assumptions for estimating principal causal effects. These include stochastic dominance (Zhang et al., 2008), moment constraints (Jo, 2002), and principal ignorability (Jo & Stuart, 2009).
8. Here, the use of the term ''local'' refers to the fact the treatment effect of interest is local to the compliers.
9. These strategies include constructing nonparametric bounds (e.g., Cheng & Small, 2006; Grilli & Mealli, 2007; Jo & Vinokur, 2011), using auxiliary information such as proper priors (Hirano et al., 2000), covariates (Jo, 2002), or secondary outcomes (Mealli & Pacini, 2013), or explicitly changing sensitivity parameters (Roy, Hogan, & Marcus, 2008).

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, *72*, 1–19.

Abdulkadiroglu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *The Quarterly Journal of Economics*, *126*, 699–748.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association*, *98*, 299–323.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, *8*, 225–246.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., & Bos, J. M. (1997). The benefits and costs of JTPA Title II-A programs. Key findings from the National Job Training Partnership Act Study. *Journal of Human Resources*, *32*, 549–576.

Cheng, J., & Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*, 815–836.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Elliott, M. R., Raghunathan, T. E., & Li, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics*, *11*, 353–372.

Feller, A., Grindal, T., Miratrix, L., & Page, L. C. (2014). *Compared to what? Variation in the impacts of Head Start by alternative child-care setting*. Retrieved from http://ssrn.com/abstract=2534811.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 21–29.

Frumento, P., Mealli, F., Pacini, B., & Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, *107*, 450–466.

Grilli, L., & Mealli, F. (2007). Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. *Journal of Educational and Behavioral Statistics*, *31*, 111–130.

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, *1*, 69–88.

Howell, W. G., Wolf, P. J., Campbell, D. E., & Peterson, P. E. (2002). School vouchers and academic performance: Results from three randomized field trials. *Journal of Policy Analysis and Management*, *21*, 191–217.

Imai, K., Jo, B., & Stuart, E. A. (2011). Commentary: Using potential outcomes to understand causal mediation analysis. *Multivariate Behavioral Research*, *46*, 861–873.

Imbens, G. W., & Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, *64*, 555–574.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge, England: Cambridge University Press.

Jin, H., & Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, *103*, 101–111.

Jin, H., & Rubin, D. B. (2009). Public schools versus private schools: Causal inference with partial compliance. *Journal of Educational and Behavioral Statistics*, *34*, 24–45.

Jo, B. (2002). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods*, *7*, 178.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, *13*, 314.

Jo, B., & Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, *28*, 2857–2875.

Jo, B., & Vinokur, A. D. (2011). Sensitivity analysis and bounding of causal effects with alternative identifying assumptions. *Journal of Educational and Behavioral Statistics*, *36*, 415–440.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, *76*, 1071–1102.

McConnell, S., Stuart, E. A., & Devaney, B. (2008). The truncation-by-death problem: What to do in an experimental evaluation when the outcome is not always defined. *Evaluation Review*, *32*, 157–186.

Mealli, F., & Mattei, A. (2012). A refreshing account of principal stratification. *The International Journal of Biostatistics*, *8*, 1–37.

Mealli, F., & Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, *108*, 1120–1131.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide: Statistical analysis with latent variables: User's guide* (6th ed.). Los Angeles, CA: Author.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles (with discussion). *Statistical Science*, *5*, 465–480.

Orr, L. L. (1999). *Social experiments: Evaluating public programs with experimental methods*. Thousand Oaks, CA: Sage.

Page, L. C. (2012). Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness*, *5*, 215–244.

Pearl, J. (2011). Principal stratification—A goal or a tool? *The International Journal of Biostatistics*, *7*, 1–13.

Peck, L. R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation*, *24*, 157–187.

Peck, L. R. (2013). On analysis of symmetrically predicted endogenous subgroups part one of a method note in three parts. *American Journal of Evaluation*, *34*, 225–236.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York, NY: Springer.

Roy, J., Hogan, J. W., & Marcus, B. H. (2008). Principal stratification with predictors of compliance for randomized trials with 2 active treatments. *Biostatistics*, *9*, 277–289.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, *7*, 34–58.

Rubin, D. B. (1980). Discussion of 'Randomization analysis of experimental data in the Fisher randomization test' by Basu. *Journal of the American Statistical Association*, *75*, 591–593.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*, 472–480.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, *31*, 161–170.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*, 322–331.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death. *Statistical Science*, *21*, 299–309.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014-4017). Washington, DC: National Center for Education Evaluation and Regional Assistance.

Stuart, E. A., Perry, D. F., Le, H. N., & Ialongo, N. S. (2008). Estimating intervention effects of prevention programs: Accounting for noncompliance. *Prevention Science*, *9*, 288–298.

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters*, *78*, 2957–2962.

VanderWeele, T. J. (2011). Principal stratification—Uses and limitations. *The International Journal of Biostatistics*, *7*, 1–14.

Zhang, J. L., & Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, *28*, 353–368.

Zhang, J. L., Rubin, D. B., & Mealli, F. (2008). Evaluating the effects of job training programs on wages through principal stratification. *Advances in Econometrics*, *21*, 117–145.

Zigler, C. M., & Belin, T. R. (2012). A Bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. *Biometrics*, *68*, 922–932.