Published in final edited form as:

J Res Educ Eff. 2012 July ; 5(3): 245–249. doi:10.1080/19345747.2012.688412.

Comments: Should Principal Stratification Be Used to Study Mediation Processes?

Tyler J. VanderWeele

Harvard School of Public Health

Principal stratification provides an approach to study the effect of an exposure on an outcome within strata defined by the effect of the exposure on some third post-treatment variable (Frangakis and Rubin, 2002). It has been used to give insight into randomized trials with non-compliance (Angrist et al., 1996), to settings where an outcome may be censored by death (Rubin, 2006), and to the study of the effects of a vaccine on post-infection outcomes (Hudgens and Halloran, 2006). There has been more recent interest in using principal stratification to study the extent to which the effect of an exposure on an outcome is mediated by an intermediate variable (e.g. Gallop et al., 2009). In this issue of the *Journal of Research on Educational Effectiveness*, Page (2012) provides evidence that the Career Academies program had substantial effect on subsequent earnings for those for whom the program would change exposure to the world-of-work, but not for those for whom it would not change exposure to the world-of-work. What I would like to discuss in this commentary is the question of what we learn about mediation from these principal stratification analyses. I believe that principal stratification analyses in the context of an intermediate can be helpful or illuminating. I also think the results of Page (2012) concerning the Career Academies are of substantive interest. However, as I will argue in this commentary, I do not think that the causal effects defined by the principal stratification framework correspond to mediation. Mediation as traditionally understood corresponds to the phenomenon in which a treatment or exposure changes a mediator and that change in the mediator goes on to change the outcome. This is not what the principal stratification framework captures.

Point 1. There is nothing in the principal stratification framework that corresponds to a mediated effect

In the notation of Page (2012), let Z denote the treatment, M the intermediate, and Y the outcome. Let M_0 and M_1 denote the potential values of the intermediate, and Y_0 and Y_1 the potential values of the outcome, under the control condition and the treatment condition respectively. For simplicity, let us assume that the intermediate is binary.¹ Within the principal stratification framework, we might consider the effect of treatment on the outcome for the subgroup for whom M_0 and M_1 are both 0; this would be $E[Y_1 - Y_0 | M_0 = M_1 = 0]$. Or we might consider the effect of treatment on the outcome for the subgroup for whom M_0 and M_1 are both 1; this would be $E[Y_1 - Y_0 | M_0 = M_1 = 1]$. These two effects are

¹In the actual analysis of Page (2012) the mediator is considered as a variable with three levels; all of the points in this commentary are equally applicable to a mediator with three levels.

sometimes referred to as principal stratum direct effects because the treatment does not change the mediator and so all of the effect of the treatment on the outcome must be direct.

Alternatively, we might consider the effect of treatment on the outcome when the intermediate is 0 without treatment and 1 with treatment; this would be $E[Y_1 - Y_0 \mid M_0 = 0, M_1 = 1]$. This is sometimes referred to as an associative principal stratum effect (Frangakis and Rubin, 2002). There may be a temptation to take this associative effect as a mediated effect. However, this would not be correct. As Page (2012) points out, this associative effect captures the *total* effect of the treatment on the outcome (i.e. both the direct and the indirect effects) for the subgroup for which the treatment changes the value of the intermediate from 0 to 1. Because the associative effect captures both the direct and indirect effect for this subgroup, the associative effect can be non-zero even if the intermediate has no effect on the outcome i.e. even if M is not a mediator. There simply is nothing in the principal stratification framework that corresponds to a mediated effect.

Point 2. We may have zero principal strata direct effects and a non-zero associative effect but no mediation

Consider Figure 1. Suppose that M has no effect on Y whatsoever so that it is not a mediator. Although treatment is randomized, the mediator is not and so we may have an unmeasured common cause of the intermediate and the outcome. In the analysis of Page (2012), the variable U might be baseline pre-treatment motivation. Because of this unmeasured common cause, we may have that the outcome is low whenever the intermediate is low, and that the outcome is high whenever the intermediate is high. For example, suppose that students with low motivation ($U=0$) are unresponsive to treatment both in terms of exposure to the world-of-work (either $M_0=M_1=0$ or $M_0=M_1=1$) and in terms of earnings ($Y_0=Y_1$). In this case, both of the principal strata direct effects will be zero. Suppose that students with high motivation ($U=1$) are responsive to treatment, both in terms of exposure to the world-of-work (so that $M_0=0, M_1=1$) and in terms of earnings ($Y_1 > Y_0$); then the associative principal stratum effect $E[Y_1 - Y_0 \mid M_0=0, M_1=1]$ will be non-zero. We may have zero principal strata direct effects, and a non-zero associative principal strata effect, as in Page (2012), even if exposure to the world-of-work has no effect on earnings and thus does not serve as a mediator at all. This may occur because it may be only the high motivation students for whom their earnings are responsive to treatment even if all of this effect is direct (cf. VanderWeele, 2011).²

Point 3. Analyses of effects within principal strata are more akin to subgroup analyses than to mediation analyses

Social scientists generally distinguish between “mediators” and “moderators”. A mediator is a variable on the pathway from the treatment to the outcome. A moderator is a variable such that the effect of the treatment on the outcome differs for different levels of the moderator variable. In a randomized trial, for instance, we might estimate the effect of the treatment on

²I am not arguing here that exposure to the world-of-work is not a mediator, only that the analyses of Page (2012) are perfectly consistent with it not being a mediator.

the outcome within strata of baseline pre-treatment covariates. This would be a moderator analysis. Within the principal stratification framework, the principal strata are effectively viewed as pre-treatment characteristics of an individual (Frangakis and Rubin, 2002). Assessing the effect of treatment on the outcome across different principal strata is essentially looking at whether the effect of treatment – the total effect – is different for different subgroups. In the terminology of social scientists, this would be a “moderator” analysis. Effects within principal strata are akin to moderator or subgroup analyses, not mediation analyses.

Point 4. For the purposes of mediation, principal stratification moves the problematic assumption referred to as “ignorability” from what is necessary for identification to what is necessary for interpretation

Page (2012) points out that by using a Bayesian approach to principal stratification and by using covariates to attempt to predict which principal stratum different individuals belong to, she is able to avoid the “ignorability” assumption required in many analyses of direct and indirect effects.³ This assumption is essentially that, within strata of the covariates, the intermediate is effectively randomized. Conceived of another way, the assumption is that we have been able to control for all of the common causes of the mediator and the outcome. Page (2012) does not make this assumption in her analysis. However, as has already been suggested above, at least for the purposes of mediation, the assumption seems to crop up again in the interpretation of the principal strata effects. If we have not controlled for potential common causes of the intermediate and the outcome then we have not ruled out the possibility that an unmeasured common causes is responsible for the phenomenon that the treatment has an effect on the outcome only when the treatment has an effect on the intermediate. As noted above, pre-treatment motivation might alter both the effect of treatment on exposure to the world-of-work and the effect of treatment on earnings. Page (2012) draws inferences on the principal strata effects without making an ignorability assumption, but she does not rule out an unmeasured intermediate-outcome common cause as a possible explanation for her results.

Point 5. Principal stratification tells us more about surrogacy than mediation

What then do we learn from an analysis such as that of Page (2012)? We learn that the Career Academies program had substantial effect on subsequent earning only for those for whom the program changed exposure to the world-of-work. In my view, this itself is of interest. It effectively tells us that exposure to the world-of-work is a useful surrogate for the effect of the Career Academies program on subsequent earnings. It may, as Page (2012) points out, also indicate that the program affects earnings *through* exposure to the world-of-work; but it may not. The analyses of Page (2012) are perfectly consistent with exposure to

³Although Page's (2012) analysis does not make this ignorability assumption, the assumptions she does make do not suffice to point-identify her estimates; this means that as the sample size tends to infinity the length of her posterior intervals will not shrink to zero; they will not converge to a single point. To point identify principal strata effects much stronger assumptions, such as those made for direct and indirect effects, are needed.

the world-of-work having no effect on subsequent earnings and with exposure to the world-of-work not being a mediator at all. From the principal stratification analyses, we learn about surrogacy but not about mediation.

In my view, a clear distinction should be drawn between surrogacy and mediation. Certainly there are elements of surrogacy and mediation that are similar. Both consider the effect of a treatment on an outcome and how this effect relates to some post-treatment variable. Historically, the statistical methods that have often been employed in these two settings have been similar. However, surrogacy concerns whether the effect of a treatment on an outcome can be predicted by the effect of a treatment on an intermediate, whereas mediation concerns whether the effect of treatment on an outcome is *through* the intermediate. A good surrogate may often be a mediator, but it need not be. As suggested by Figure 1 and the discussion above, we may have a good surrogate (i.e. one that predicts the effect of treatment on the outcome) even if the surrogate has no effect on the outcome at all.

In terms of potential outcomes, surrogacy and mediation correspond to different formal causal effects. The associative effect from the principal stratification framework, captures the effect of the treatment on the outcome when the treatment changes the intermediate, i.e. it captures surrogacy. And, ideally, for a good surrogate, the effect of the treatment on the outcome will be zero whenever the effect of the treatment on the surrogate is zero (Frangakis and Rubin, 2002). This was exactly what was found by Page (2012). However, for mediation, the concept of a “natural indirect effect” (Pearl, 2001) is the appropriate concept. The natural indirect effect captures whether the treatment changes the intermediate and whether that change in the intermediate changes the outcome and this is what we generally mean by mediation. Methods for natural direct and indirect effects are described elsewhere (Imai et al., 2010; Valeri and VanderWeele, 2012) and effectively generalize the direct and indirect effect concepts traditionally employed by social scientists.

Summary

In summary, then, I think that principal stratification analyses such as that of Page (2012) can be of interest, and I think her substantive conclusions are of interest. But, if an investigator is specifically interested in studying mediation, then I do not think that principal stratification provides the right approach to do so. Methods for natural direct and indirect effects, not principal stratification, should be used to study mediational processes. Mediation and surrogacy are not equivalent. The approach, definitions, and methods should match the questions of substantive interest.

References

- Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*. 1996; 91:444–472.
- Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002; 58:21–29. [PubMed: 11890317]
- Gallop R, Small DS, Lin JY, Elliott MR, Joffe M, Ten Have TR. Mediation analysis with principal stratification. *Statistics in Medicine*. 2009; 28:1108–1130. [PubMed: 19184975]
- Hudgens MG, Halloran ME. Causal vaccine effects on binary post-infection outcomes. *Journal of the American Statistical Association*. 2006; 101:51–64. [PubMed: 19096723]

- Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychological Methods*. 2010; 15:309–334. [PubMed: 20954780]
- Pag LC. Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness*. 2012 in press.
- Pearl, J. *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*. San Francisco: Morgan Kaufmann; 2001. Direct and indirect effects; p. 411–420.
- Rubin DB. Causal inference through potential outcomes and principal stratification: application to studies with “censoring” due to death (with discussion). *Statistical Science*. 2006; 21:299–321.
- Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Revised for *Psychological Methods*.
- VanderWeele TJ. Principal stratification: uses and limitations. *International Journal of Biostatistics*. 2011; 7:1–14. Article 28.

Figure 1.
Associative Principal Strata Effect with No Mediation: U – pretreatment motivation; Z – treatment; M – exposure to the world-of-work; Y – earnings.