MULTIVARIATE
BEHAVIORAL
RESEARCH
The Journal of the Society of
Multivariate Experimental Psychology

Multivariate Behavioral Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

Introduction to the Special Issue on Propensity Score Methods in Behavioral Research

Robert M. Pruzek [a]

[a] University of New York, Albany

Available online: 09 Jun 2011

To cite this article: Robert M. Pruzek (2011): Introduction to the Special Issue on Propensity Score Methods in Behavioral Research, Multivariate Behavioral Research, 46:3, 389-398

To link to this article: http://dx.doi.org/10.1080/00273171.2011.576618

Multivariate Behavioral Research, 46:389–398, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2011.576618

Psychology Press
Taylor & Francis Group

Introduction to the Special Issue on Propensity Score Methods in Behavioral Research

Robert M. Pruzek
University of New York, Albany

This issue includes six articles that present logic, methods, and models for causal analyses of observational data, in particular those based on propensity score (PS) methods. The articles include a general introduction to propensity score analysis (PSA), uses of PSA in mediation studies, issues involved in choosing covariates, challenges that often arise in PSA applications, hierarchical data issues and models, and an application in an educational testing context. In this editorial I briefly summarize each article and make a few recommendations that relate to future applications in this field: the first pertains to how propensity score (PS) work could profit by connecting it with stronger forms of randomized experiments, not just simple randomization; the second to how and why graphical methods could be used to greater advantage in PSA studies; then why it might be helpful to reconsider the meaning of the term "treatments" in observational studies and why conventional usage might be modified; and finally, to the distinction between retrospective and prospective approaches to observational study design, noting the advantages, when feasible, of the latter approach.

PS theory and methods have been on the scene for more than a quarter of a century now and have become central to both theoretical statistical inquiry and to many fields of applied science. In the 10 years between 1997 and 2007 the numbers of published applications of PSA in the health and medical sciences grew from about 15 articles per year to more than 300 per year (retrieved from

Correspondence concerning this article should be addressed to Robert M. Pruzek, University of New York at Albany, Department of Educational and Counseling Psychology, The University at Albany, SUNY, Educ. 233, 1400 Washington Avenue, Albany, NY 12222. E-mail: rpruzek@uamail.albany.edu

http://www.ncbi.nlm.nih.gov/pubmed). Roughly, the increases have followed an exponential curve that shows no sign of tapering off, as more than 2,000 articles that concern PSA have now been published in the health sciences. In contrast, as shown in a recent article by Thommes and Kim (2011), the threshold of 15 articles per year was not reached until about 2005 in the combined fields of psychology and education and only reached 30 articles per year by 2009. That PS studies are beginning to appear in social, educational, and behavioral sciences, and seem often to yield useful results, suggests that the relative dearth of applications so far in these fields has more to do with the insularity of distinctive research domains than with general unsuitability of PS methods in behavioral research.

The paper that started the upsurge of work on PS methodology was written by Rosenbaum, a student of Rubin, jointly with Rubin (Rosenbaum & Rubin, 1983). This paper defined propensity scores, proved a number of fundamental theorems about them, and discussed their potential uses. Numerous statisticians, both theoretical and applied, quickly saw the potential of the basic ideas, with the result that a veritable cottage industry of PS activity was soon stimulated. Just as important, both coauthors of that paper, and their students, have continued to contribute to the literature of this field, with numerous articles and books (see Rosenbaum, 2002, 2010; Rubin, 2006). Other important work is in development. (Indeed two of those students, Hill and Stuart, are cited at several points later and are authors of articles that appear in this issue.)

This Special Issue should be seen as an attempt to broaden awareness among behavioral researchers to PSA methods, to provide background and resources that can facilitate understanding of the methodology, and to report new thinking and methods that seem to hold promise for advancing PS applications in behavioral research. In the section that follows I provide some brief comments on each of the respective articles, partly to draw attention to the distinctive concepts and terminology of PS research and also to help put the articles in perspective, both in relation to one another and in relation to key literature of this field. In the final section I offer suggestions for further learning about these and related topics as well as some thoughts about how PS methodology might be enhanced or elaborated to advance the usefulness, efficiency, or comprehensiveness of PS methods in behavioral research.

BRIEF COMMENTS ON THE SPECIAL ISSUE ARTICLES

The first article by Austin draws heavily on his own long and steady stream of contributions to the propensity score literature to describe and discuss the basics of PS methods; he also includes reference to his recent tutorial on this topic

(Austin, 2011) that was written largely as a companion to the current article. Austin reviews the key features of the potential outcomes (sometimes referred to as counterfactual) framework, but his focus is on concepts underpinning these methods. Special attention is given to covariate balance, which is central to nearly all applications of the basic methods. He distinguishes four ways of using estimated propensity scores to estimate treatment effects, including matching, stratification, inverse probability weighting, and related regression methods. Austin discusses advantages of PS methods over conventional regression methods in observational studies. He also notes the differences between estimating the average treatment effect (ATE) and the average effect for the treated (ATT). His ultimate goal is to introduce the reader to the basic idea of the PS and to describe how methods based on it can be used to reduce or eliminate the effects of confounding when using observational data to estimate treatment effects.

In the second article, Jo, Stuart, MacKinnon, and Vinokur focus on the use of propensity scores in mediation analyses, that is, studies that consider how variables intermediate to an independent variable and a response can mediate the response. They note that standard practice in mediation analysis entails reliance on untested distributional and functional form assumptions, whereas PS methods offer some notable alternatives. They use propensity scores to compare individuals in treatment and control groups who would have had the same value of the mediator had they been assigned to the same treatment condition. The authors pay particular attention to identifying underlying assumptions and point out possible advantages of propensity scores in the context of mediation analyses, using alternative assumptions, compared with what has become orthodox in mediation research. They illustrate their PS approach in an example where the goal is to reduce depression in a job search interventional context. They found statistical differences for individuals who would have an improved sense of mastery when in the treatment condition but no discernable differences for persons who would not have improved their sense of mastery under treatment.

Kelcey, in the third article, addresses the fundamental question of how appropriately to select covariates for use in constructing propensity scores. His article contains a lengthy review of the relevant literature as well as careful analyses of the merits and demerits of different ways to go about covariate selection. Kelcey discusses the central role of the strong ignorability assumption in PSA and notes that there is sometimes a trade-off between maximizing balance and reducing bias of PS estimates; he then analyzes various implications of different courses of action in addressing this trade-off. He notes that most PS applications have followed the general principle advocated by Rubin and his colleagues to use all available covariates that appear related to both "exposure" (the treatment/control distinction) and the outcome when constructing propensity scores. Yet, as he notes, propensity scores are generally constructed without explicit reference to responses. Nevertheless, given that the initial covariates should theoretically be "related" both to responses and exposure, one might like to have more than just subject matter theory to facilitate covariate selection. He uses a simulation study to explore whether, or to what extent, covariate selection might be improved through use of information available in pretreatment response variable proxies to approximate covariate-outcome relationships. Several issues raised by this author are also considered by Austin is his discussion of covariate selection.

In the fourth article, Hill, Weiss, and Zhai tackle problems that have come to be seen as especially challenging in many PS studies, particularly problems concerning choices in PS estimation strategies, matching and weighting implementation strategies, and choice of balance diagnostics and final analysis models. They show that PS estimates can vary widely from one another with different combinations of these choices, which naturally raises questions about how to select from among the possibilities. They also highlight several fundamental concepts, including covariate balance and, like Kelcey, the role of strong ignorability in PS applications. Like Austin, they note that the logic of randomized experiments is connected to the use of propensity scores in observational studies. Their contribution is worthy of special study because they not only discuss challenges that have often plagued applications but also explore a new line of constructive alternatives for causal analysis, one that does *not* need propensity scores. They apply and discuss this new method using real data, an example from education concerning the issue of holding children back in early elementary school to show how one can adjust for the effects of hundreds of covariates, in their case 256 of them (these being pared from a larger set of 500!). They use a methodology based on Bayesian additive regression trees (BART). Beyond their conceptual and methodological contribution, the authors exhibit creativity in how they comprehensively present results for several methods where the number of covariates, by ordinary standards, is almost impossibly large.

Thoemmes and West, in the penultimate article, examine the problem of estimating treatment effects in observational studies when the data are clustered. They distinguish among and describe differences in assumptions for several different model specifications that might be used for constructing propensity scores, including single level and fixed effects models and two random effects models. They note that multilevel modeling methods can be employed either for cases where the clusters are central features of a design or where clusters should be taken into account even when they are incidental to the central purposes of a study. As in all other articles in this issue, the authors restrict attention to binary treatment comparisons, such as when a treatment or condition or behavior is compared with a control. They then compare different analysis models for both simulated and real data. Their real data problem is similar to that of Hill, Weiss, and Zhai in that both examples concern the effects of holding schoolchildren back a grade, but it is different in one fundamental way too: Thoemmes and West use data based on a prospective design, not a retrospective one (see later).

In the sixth and final article, Lottridge, Nicewander, and Mitzel compare online tests with paper tests using PS methods. Their article is the only one in this issue for which the primary focus is on a subject-matter question, not methods. They review the role of computer delivery of tests and how that has increased in the past few years; the authors identify perceived advantages to include greater flexibility of scheduling, easier tailoring to specific needs, and more rapid scoring and reporting. Still, paper test formats have a much longer history and are more familiar to school personnel and students. The authors use two approaches to compare both the score and construct equivalence of these two formats. End-of-year Algebra and English tests were compared in the context of a statewide testing program in Grades 8 and 9. PS matching served as the primary vehicle for their comparison. Results of their analyses suggest that the two formats yield scores with about the same level of reliability and with about the same correlations with external measures after covariate-based adjustments using PS matching.

FURTHER THOUGHTS RELATED TO FUTURE STUDIES

Anyone aiming to learn more about modern methods for causal analyses based on observational data currently has numerous articles and books to study. For those who have not yet examined the PS literature, several of the books and articles that the authors cite can serve as helpful introductions (see especially Austin, 2011; Morgan & Winship, 2007; Rosenbaum, 2002, 2010; Stuart, 2010; for an expository article, see Rubin, 1997). Recent articles of special relevance include Cook and Steiner (2010), Shadish (2010), and West and Thoemmes (2010), all of which appear in a special section of *Psychological Methods* focused on relationships between Campbell's and Rubin's methods as they pertain to causal studies. Harder, Stuart, and Anthony (2010) focus on covariate balance in the context of assessing causal associations. Stuart and Ialongo (2010) examine matching for selection of participants who are to be followed up. All of the latter articles appear in behavioral or psychological methods journals, and as such, deal with issues of special relevance to social scientific researchers.

Beyond the preceding books and articles, there is now a good deal of software that can facilitate execution of PS analyses, and more. Software documentation can help a researcher to understand the basic methodology. The software platform R, which is both free and especially well suited to PS applications, is notable because it provides numerous basic and refined packages that can support PS studies in both their numerical and graphical aspects. The *Journal of Statistical Software*, and certain others as well, are good sources for extended documentation of many of these packages where it can be seen that both numerical and graphical PS methods have a strong presence in this literature.

There are four topics I would like briefly to discuss. The first concerns the connection between randomized controlled trials (RCTs), which Rubin has taken as his starting point in the development of his potential outcomes model, and current observational study practice. Second, I will comment on the role of graphics and visualization in PS contexts. My third topic has to do with relationships between definitions or meanings of the term *treatments* in observational studies compared with RCTs. The fourth has to do with a distinction between retrospective and prospective approaches to observational study design.

When RCTs are feasible, that is, when they are administratively possible and ethical, they are generally to be preferred to observational studies for the simple reason that all covariates, observed or not, are accounted for (on average) by the act of randomizing units to treatments. It is commonly said that RCTs represent the "gold standard" for treatment comparison when the aim is to interpret effects as causal. Although this logic is broadly accepted, there is reason to dig deeper. Specialists in experimental design routinely recommend designs that take individual differences into account prior to randomization, especially using various blocking methods. In fact, design specialists often argue against simple randomization that ignores individual differences on the grounds that simple designs are not only less efficient compared with blocked designs (sometimes hugely so) but also that restriction to simple randomization generally weakens or even denies the investigator's ability to discover interactions between covariates and treatments when they exist. If RCTs are to serve as models for observational studies, then it seems only reasonable to suggest that especially efficient and effective forms of experimental design methodology should be more directly connected to their observational counterparts.

Given the extensive literature of observational methods research, it should not be surprising that some scholars have considered closely related issues (cf. Rosenbaum, 1991), but to my knowledge there are few applied causal studies (at least in the behavioral sciences) where the role of individual differences is taken seriously at the point of response data analysis. There are of course basic difficulties in connecting observational methodology to, say, randomized block designs (after all, randomized assignments in RCTs are done after taking individual differences into account), yet it would appear that concepts and general thinking associated with block designs can serve an important role to help spell out questions for study as well as ways of collecting and analyzing data. For example, one might use covariate information, perhaps based on initial clustering, to form subsets of data for display and analysis and proceed to use these in the context of matching or stratification. Focusing more explicitly and comprehensively on the possibilities of *interactions* between covariates and treatments would constitute an interesting change in standard analytic practice.

Rosenbaum's (1991) work, among others, on optimal designs also suggests that efficiency gains might be expected to follow from subsetting and related methods at the time of analysis. By tightening the connections between observational studies and randomized block approaches to experiments, I believe investigators could gain traction at a conceptual level and when strategizing comprehensive approaches to observational data analyses.

The second topic I want to mention concerns the use of graphics and visualization of data in PS analyses—and in observational analyses more generally. It is rare to find examples of modern graphics even in recent books that deal with causal analysis, and the constraints of many journals are such as to make it difficult to publish extended graphics except in special cases. Yet, numerical summaries can nearly always be complemented by use of effective (modern) graphics, often to great advantage. This is because graphics can help one learn details of what data have to say and can lead to new insights and new analyses that can go well beyond one's initial questions. Many analysts have come to see modern graphics as central to their applied research and it is not difficult to give examples in which visualizations of data have played a key role. Despite the dearth of examples in causal research articles and textbooks, I think this issue is too important to ignore as research goes forward in this field.

Stuart (2010) provides a useful summary of software sources in different packages. Helmreich and Pruzek (2009) document several graphical methods that were designed specifically for PSA work; they consider ways to assess balance, to display outcomes following stratification, and generally to complement PSA summary statistics. Pruzek and Helmreich (2009) concentrate more specifically on uses for graphics, such as in matching, and illustrate how they can provide special value in the analysis of randomized block designs, such as exposing interactions, outliers, and patterns in data. The work of Sarkar (2008) on lattice graphics (in R) could also play a major role in clarifying results when one might like to focus on details of what data may have to say.

My third issue concerns the possibility that there may be notable differences between *treatments defined in the context of RCTs* and *treatments defined by how individuals choose to act or behave.* In the observational research methods literature, where mathematical thinking has been dominant, treatments are generally defined in ways that suggest no conceptual difference between so-called observational treatments and their RCT counterparts. That is, when two groups are compared, one is usually called a "treatment," the other a "control." But I believe there may be good reasons in many observational studies to question the veracity of any strict categorization based on two categories, even if one's basic interest lies with what, generically, might be called "two treatments." When individuals choose to behave in particular ways, or "select" their own treatments, then analysts may need to study closely the specific behaviors persons have chosen (that the analyst may generally call "treatments" but might better be called "classes of treatments"). The basic issue also may concern counterpart comparison "controls," which of course are often just alternative (classes of) treatments. In general, this is not just a matter of treatment compliance or adherence, common terms in Rubin's framework, because there, the only options are adhere, or not. The conceptual model for experimental science, by its usual definition, assumes that (treatment) "control" means comprehensive control. In observational frameworks, where control is not part of the picture, human action or behavior may be less easy to categorize than is implied by the words "treatments" and "controls." Therefore, I believe that variations in behavior should often be expected and the ranges of manifest behaviors carefully defined across individuals when their actions define the categories. The point I am making may be transparent to behavioral scientists, but because most applications of PS methods have occurred outside this field, the issue seems not to have received attention in causal behavioral work. How variations in behavior might play out in observational analyses might itself deserve focused study in future research.

The final topic I see to be worthy of more consideration concerns the basic difference between retrospective and prospective approaches to data collection. The retrospective approach, based on archival data, is by far the most dominant one in observational studies (as is the case for all but one of the examples in this issue, as noted earlier). But prospective approaches to data collection can offer possibilities rarely available using conventional approaches. When a prospective modality is employed one begins with a design plan that can help ensure collection of relevant data, including covariates; this approach can help define or control of key features of treatments and choice of respondents. In the medical and health sciences, where PS studies have been most common, archival data are often extensive, and data choices may be substantial. But for many (behavioral) research questions, archival data, if they are available at all, may be far from what is needed to meet the ignorability goals of observational research methods because one is at the mercy of the quality and quantity of available data. In short, the design feature of prospective observational studies seems likely to lead to better data to help ensure minimal selection bias and more informative conclusions.

CONCLUSION

This Special Issue provides only one of many looks into the methodology of PSA. These articles, or the work on which these authors build, can provide assistance to nearly any behavioral researcher who might consider work in this field. But several strands of work on causal studies have of necessity been left out too. Not only is there a literature deriving from work of statisticians and methodologists, which is dominant here, but there is also a larger and older literature of causal studies to which PS methodology has interesting connections. I want to point to just two areas that seem to me to be especially worthy of consideration.

One valuable class of work was briefly noted by Hill, Weiss, and Zhai (this issue), namely, the work of Judea Pearl. Pearl (2010) summarizes his major concepts and methods and demonstrates that he has made major contributions to thinking and methods for causal analyses. His basic approach, which is strongly connected to structural equation modeling, is fundamental as well as comprehensive; his recent review (Pearl, 2009) should also be noted.

Ni Bhrolch'ain and Dyson (2007) discuss several aspects of causation in demography; they focus mostly on birth and death rates as well as fertility. But the key value of their article concerns guidelines one might consider to help establish cause. They provide a useful table listing 10 guidelines, 3 of which are essential: time ordering, evidence of mechanism, and uniqueness of interpretation. Like so many scholars of causal analysis, they urge caution in making causal inferences, particularly in situations where there may be notable policy implications.

Finally, I would be remiss if I did not thank the authors and several reviewers who provided major assistance in helping to put this Special Issue together. Several authors also served as reviewers of manuscripts of others and made many suggestions that were essential to this project. I join with the editorial staff to thank you all.

REFERENCES

Austin, P. C. (2011). A tutorial and case study in propensity score analysis: An application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behavioral Research, 46,* 119–151.

Cook, T. D., & Steiner, P. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods, 15,* 56–68.

Harder, V. S., Stuart, E. A., & Anthony, J. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15,* 234–249.

Helmreich, J. E., & Pruzek, R. M. (2009). PSAgraphics: An R package to support propensity score analysis. *Journal of Statistical Software, 29*(6). Retrieved from http://www.jstatsoft.org/v29/i06 paper

Morgan, S., & Winship, C. (2007). *Counterfactuals and causal inference*. New York, NY: Cambridge University Press.

Ni Bhrolch'ain, M., & Dyson, T. (2007). On causation in demography: Issues and illustrations. *Population and Development Review, 33,* 1–36.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3,* 96–146.

Pearl, J. (2010). *Causality: models, reasoning, and inference* (2nd ed.). Cambridge, UK: Cambridge University Press.

Pruzek, R. M., & Helmreich, J. (2009). Enhancing dependent sample analyses with graphics. *Journal of Statistics Education, 17*(1). Retrieved from http://www.amstat.org/publications/jse/v17n1/helmreich.pdf paper

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, Series B, Methodological, 53,* 597–610.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

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

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Rubin, D. B. (1997). Estimation from nonrandomized treatment comparisons using subclassification on propensity scores. *Annals of Internal Medicine, 127,* 8, 757–763.

Rubin, D. B. (2006). *Matched sampling for causal effects.* New York, NY: Cambridge University Press.

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R.* New York, NY: Springer.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods, 15,* 3–17.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25,* 1–21.

Stuart, E. A., & Ialongo, N. S. (2010). Matching methods for selection of subjects for follow-up. *Multivariate Behavioral Research, 45,* 746–765.

Thommes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46,* 90–118.

West, S. G., & Thoemmes, F. J. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods, 15,* 18–37.