Critique Papers on Causal Inferences

Tony Tan

Centre for Educational Measurement

University of Oslo

MAE4051: Selected Topics in Educational Measurement

Prof Jan-Eric Gustafsson

9 April 2021

Propensity Scores

Sullivan, A. L., & Field, S. (2013). Do preschool special education services make a difference in kindergarten reading and mathematics skills?: A propensity score weighting analysis. *Journal of School Psychology*, *51*(2), 243–260. https://doi.org/10.1016/j.jsp.2012.12.004

Summary

Sullivan and Field (2013) attempted to investigate the marginal benefit created by a special education program mandated by the US *Individuals with Disabilities Education Act* (IDEA). The 1986 amendments to the IDEA legislation imposed legal requirement on the states to create a sequence of intervention programs targeting three stages of development for children with needs: 1) from birth to 2-year-old, 2) 3-year to kindergarten entry, and 3) kindergarten to 21-year-and-11-month-old. This paper focused on the middle segment of the intervention sequence named preschool special education services and followed children longitudinally through the delivery period. The study began by measuring children's probability of being admitted into the special education program using a logit regression, and at the end of the intervention period, compared the average performance scores in maths and reading of the children who actually received the preschool special education treatment against the average scores of those who have not. Using the propensity score weighting technique, the authors interrogated the counterfactual question of "what if" the treatment were never applied and reached an conclusion that children in treatment group would have been better off academically had they not received any intervention at all. This disappointing result lent itself to an existing literature of similar negative findings evaluating special education effectiveness, cautioning the rosy objectives of preschool special education services originally marketed by policy makers.

Causal Question

Would the children who received special education services have been better off academically, on average, had they not received such services?

Validity

Internal and External Validity

Historically, internal validity referred to inferences about whether "the experimental treatments make a difference in this specific experimental instance" while external validity asked "to what populations, settings, treatment variables, and measurement variables can this effect be generalized" (Campbell & Stanley, 1963, p. 5). Cook and Campbell (1979) advanced the idea of internal validity to the question whether the covariation observed between the independent and dependent variables were resulted from a *causal* relationship, whereas external validity further asks whether such cause-effect relationship holds over certain variation in persons, settings, treatment variables, and measurement variables.

In order to support an inference that the observed covariation between A and B reflects a causal relationship, Shadish et al. (2002) prescribed a trifecta that 1) A preceded B in time, 2) A covaries with B, and 3) no other explanations for the relationship are plausible. It is too often the third strand that undermines the internal validity of inference making—the relationship between A and B is not causal because it could have occurred even in the absence of the treatment and that it could have led to the same outcomes that were observed for the treatment. Amongst the list of potential threats to internal validity (Shadish et al., 2002, pp. 54–61), maturation presents the strongest challenge to Sullivan and Field (2013). Children who have been identified as "in need" so early in life ("early onsetters") can be reasonably believed to be in possession of different developmental profiles from children who showed needs later in life ("late onsetters"). As participants in the treatment group mature, gaps in academic performance may well emerge out of such delayed developmental trajectories with or without education services. It is therefore not preschool interventions that "caused" lower academic scores but the two-tier growth profiles that did. Failing to rule out such alternative, and rather plausible, explanation weakens the internal validity of inference made by the authors.

Weak external validity has also been acknowledged by the authors in Section 4.2 of the paper. Inferences can only been drawn, first of all, over children with mild to moderate impairments resultant from the sample deletion procedure; while it were the children with the most severe impairment that policy makers wished to monitor and retain ("interaction of the causal relationship with units" by Shadish et al. (2002)). Secondly, Sullivan and Field (2013) evaluated only the academic performance of young children to the exclusion of other developmental markers such as motor-behavioural and social-affective skills—all vital policy objectives along with reading and maths scores, if not more important, for kindergarten-entry age children ("interaction of the causal relationship with outcomes"). Lastly, the averaging procedure in calculating ATT washed out important differences across race and socio-economic groups, important factors reported by prior literatures as non-ignorable ("context-dependent mediation").

Construct Validity

Construct validity concerns itself with the degree of agreement between the concept the researchers intended to understand (e.g., academic performance) and the procedure as well as instrument they employed to capture and measure such concept (e.g., sum scores in maths and reading tests). Amongst the various threats proposed by Shadish et al. (2002, pp. 72–81), Sullivan and Field (2013) were particularly susceptible to "inadequate explication of constructs" and "construct confounding". The concept of academic performance can be thought as the end result of a sequence of social activities: academic input (I)–academic processing(P)–academic output(O). When academic scores were low, one is unable to ascertain whether it was the result of inferior teaching (xPO), lack of learning skills (IxO), or inability to demonstrate or document learning outcome to observers (IPx). Although both xPO and IPx may show up as low academic scores, the "causal pathways" cannot be more different—a situation not assisted by the ECLS-B early reading and math batteries used in Sullivan and Field (2013) since none was designed to locate the source(s) of academic deficiency.

Construct confounding occurs when the concept under investigation has not been careful separated from other related concepts. Sullivan and Field (2013) clearly wished to study "academic performance" of young children but such construct covary particularly strongly for this age group with attention span and sociability. It is not unreasonable to conjecture that recipients of the special education program may not develop the above-mentioned skills at the same pace as their counterparts. Effectively, the ECLS-B batteries employed by Sullivan and Field (2013) were capturing young children's short attention spans and under-developed social skills and presenting them as inferior academic performance. Both inadequate explication of constructs and construct confounding have, therefore, weakened this study's construct validity, undermining its inference of "special education causing lower academic performance".

Statistical Conclusion Validity

By Cook and Campbell (1979), statistical conclusion validity refers to the appropriateness of statistical techniques employed by the researcher for the purposes of inferring whether the presumed independent and dependent variables indeed covary. The propensity score weighting technique employed by Sullivan and Field (2013) successfully circumvented many pitfalls summarised by Shadish et al. (2002, pp. 45–52) except for the "restriction of range" threat to statistical conclusion validity. Due to the necessity of constructing a region of common support, children are purposefully excluded from analyses if their probabilities of being accepted into the special education program fall outside of the 1% to 82% range. This practice is especially concerning for the above-82% group since these are the young children with demonstrated need for urgent education assistance. Under the law of diminishing marginal returns, it is more than probable that it is this most-in-need group that would have responsed best and most rapidly to special education interventions. The wholesale omission of this positive outcome pool may have well contributed to the underestimation of the project effectiveness.

Appropriateness of Methods

Sullivan and Field (2013) largely followed the propensity score anlysis procedure prescribed by Imbens and Rubin (2015) and Imbens (2015) in assessing causal effects. At the first stage DESIGN, the authors established sufficient overlapping by discarding some units from the original sample in order to establish the region of common support; the second stage SUPPLEMENTARY ANALYSIS, however, appeared to be lacking in Sullivan and Field (2013) where the plausibility of unconfoundedness shall be further addressed through pseudo-average treatment effect on the pseudo-outcome for trimmed sample (see Imbens, 2015, pp. 383–384); such absence would cast doubt on any result in the third stage ANALYSIS over the source of average treatment effect.

One highlight on methodology is the Bayesian approach to AAT sampling weights. Sullivan and Field (2013) correctly pointed out the "curse of dimensionality" when computing w_i for $D_i = 0$ cases and provided sufficient derivation through Bayes formula in reaching the form

$$w_i = \frac{\mathbb{P}\left(D = 1|z_i\right)}{1 - \mathbb{P}\left(D = 1|z_i\right)} \cdot \frac{\mathbb{P}\left(D = 0\right)}{\mathbb{P}\left(D = 1\right)}, \text{ for } D_i = 0.$$

The authors, however, stayed short of advocating for a wider adoption of this approach to resampling weights but gave in to conventional literature in order to maximise comparability. This weighting formula overcome the peculiarity of the conventional scheme (only the first term in the formula above) summing to twice the size of the treated subsample and provided a more intuitive formation of summing to the sample size. Stronger advocacy can be expected from continuing research in popularising Sullivan and Field's (2013) weighting formulation.

Conclusion

Sullivan and Field (2013) made a good attempt to apply the propensity score technique for causal assessment of preschool special education data. Despite some omissions in statistical procedure, what limited this paper's impact on policy was *not* the econometric methodology it employed but the weak inferential validity. Since "[v]alidity is a property of inferences [...] *not* a property of designs or methods," (Shadish et al., 2002, p. 34, emphasis in original text) no amount of technical sophistication is capable of compensating for validity, or the lack thereof. Sullivan and Field's (2013) result shall be interpreted narrowly based on *this* particular round of study, using *this* sub-sample to quantify *this* particular sub-set of outcome measures, subject to *these* particular restrictions, omissions and commissions, and based on *these* many statistical assumptions which may or may not have been met. A naïve interpretation of "special education *causes* even worse academic outcomes" shall be rejected out right. After all, an absence of evidence is *not* the evidence of the absence, and the total social return generated by early-life education intervention programs shall be contexualised in the general equilibrium analysis framework (e.g., through estimating the multiplier effect) rather than a partial one. Nevertheless, Sullivan and Field (2013) had made contribution to both the propensity score methodology and to the substantive debate over the direction and magnitude of the effectiveness of one social project.

Instrumental Variables

Hanandita, W., & Tampubolon, G. (2014). Does poverty reduce mental health? An instrumental variable analysis. *Social Science & Medicine*, *113*, 59–67. https://doi.org/10.1016/j.socscimed.2014.05.005

Summary

Hanandita and Tampubolon (2014) investigated the causal relationship between poverty and mental health decline using an instrumental variable (IV) approach in order to overcome the endogeneity problem. Using a sample size of 577,548 across 440 districts in Indonesia and precipitation anomaly as the IV, the authors were able to quantify the expenditure/income elasticity of mental disorders as −0.62—a result five times stronger than that of the non-IV approach and robust to various stress tests. Moreover, income inequality also appeared to carry explanatory power to mental health concerns in addition to that of poverty, suggesting both the position (quantity of income) and the shape (distribution of income) of the income curve as policy variables worth pursuing for the betterment of population mental welfare.

Causal Question

Does poverty cause poor mental health?

Validity

Since the previous critique paper has documented a typology of validity, subsequent analyses would not repeat such content but focus on the application on the paper Hanandita and Tampubolon (2014).

Statistical Conclusion and Internal Validity

Statistical conclusion validity concerns itself with whether the presumed cause and effect covary and how strongly, whereas internal validity asks whether the observed covariation is causal in nature. Hanandita and Tampubolon (2014) carefully avoided all threats to statistical conclusion validity put forward by Shadish et al. (2002, Table 2.2, p. 45). The endogeneity problem, for example, would have violated the Gauss-Markov assumption of $\mathbb{E}\left(\varepsilon_i|\boldsymbol{x}_i\right) = 0$; Hanandita and Tampubolon (2014) not only restored independent error condition through the introduction of an IV (see, for example, Greene (2018) Chapter 8 for a technical discussion of IV), but also reported the magnitude of underestimation due to such assumption violation. By ensuring the IV to be uncorrelated with mental health but highly correlated with income, this study introduced an appropriate circuit breaker to the infinite feedback loop between poverty and mental health conditions, clearly suggesting the covariation between the two had indeed been causal and the arrow of causation points from income to mental health, not the other way around.

Construct and External Validity

Both construct and external validity deals with generalisation. In addressing the construct validity, Section 4.3 of Hanandita and Tampubolon (2014) has been careful in distinguishing expenditure from income, and reported the observed deterioration in mental health condition as a response to reduction in consumption expenditure, in order to separate permanent income changes from intermittent income shock. Since this study used large dataset collected at national level, interactions of the causal relationship with both settings and outcomes can be minimum (see Table 3.2, Shadish et al., 2002, p. 87), therefore delivering strong external validity.

Appropriateness of Methods

The instrumental variable approach adopted by Hanandita and Tampubolon (2014) served their research purpose (to overcome endogeneity problem) and claim (poverty causes mental health decline) well. The last paragraph in Section 1 of the paper paid particular justification to the key assumptions behind the IV method, namely relevance condition, validity condition and exclusion restriction and revisited the suitability of these assumptions in the third paragraph of Section 6, admitting that "[t]he quality of an instrumental variabel estimation is only as good as its story" (p. 65). Although untestable, the proposal put forward by the authors that precipitation anomaly was a random assignment procedure perfectly uncorrelated with the outcome variable (mental health condition) but covaried strongly with input variables (income) due to large proportion of the Indonesian labour force being employed in a rain-dependent agricultural sector, was a convincing one.

The model building process was also appropriate. The authors ran their IV models against their baseline counterpart (i.e., models without IV); this comparison revealed a five-fold increase in the estimated effect of poverty on mental health due to the introduction of IV, incidentally revealing the magnitude of underestimation of the naïve regression approach.

Other methodological considerations also enhanced Hanandita and Tampubolon (2014)'s credibility. The authors explored both linear (linear and LPM) and non-linear (Poisson and Probit) configurations of their models to show the reported results were unlikely to be an mere artifect of the chosen functional forms. Correctional procedures such as the incorporation of sampling weights and clustering also safeguarded variance estimates. Centring of continuous variables such as log per capita household expenditure and the Gini coefficient also enhanced interpretability of the numeric results.

Conclusion

Hanandita and Tampubolon (2014) had delivered a carefully designed study to the social science community. They elevated their research enquiry from a correlational endeavour to a causal one not only to satisfy one's methodological curiosity but to provide a conclusive response to the policy choice that "if causal links between wealth and health were confirmed, society would likely benefit from more universal access to health care and redistributive economic policy. Yet, if such causal links were rebutted, resources would be better spent on influencing health knowledge, preferences, and ultimately the behavior of individuals." (Stowasser et al. (2011) as cited in Hanandita and Tampubolon (2014)). The causal evidence presented by this paper would facilitate policy actions by updating scientific believes towards the former option and contribute to the betterment of mental health project in Indonesia and developing countries at large.

Fixed-effect Regression

White, M. P., Alcock, I., Wheeler, B. W., & Depledge, M. H. (2013). Would you be happier living in a greener urban area? A fixed-effects analysis of panel data. *Psychological Science*, *24*(6), 920–928. https://doi.org/10.1177/0956797612464659

Summary

White et al. (2013)

Causal Question

Should I pretend to do econometrics while I have no idea what I am doing?

Validity

Construct Validity

Living next to green space consists of two constructs: 1. willingness to live in a green area and 2. ability to live in a green area. Assuming people are either indifferent or universally attracted to green space, what this paper picked up is actually the ability to live there, i.e., ability to afford mortgage or rent. In real estate: location location location, so the authors were not actually measuring green, but wealth.

Internal Validity

Both being able to live in green areas and happiness are the outcome of the confounding variable income/wealth.

External Validity

1 standard deviation is a huge difference. beta estimates are marginal effect, meaning at the tangent point, a little bit to the left, a little bit to the right, you can expect this much increase or decrease. This linear approximation gets lousier the farther away you move from the tangent line. 1 SD is half a world away and none the estimates can be used for such discrete changes.

Statistical Conclusion Validity

Control vars are high correlated. Coefficients do not have cetaris paribus interpretation

Appropriateness of Methods

You can throw whatever into STATA and get some output, but having a bunch of numbers in itself does not buy you legitimacy. You used STATA's panel data functionality but didn't actually run a panel analysis.

Conclusion

This paper should get burnt at the review stage. Even the author themselves admitted this paper cannot prove causal. What a waste of time and trees.

References

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Rand McNally.

Greene, W. H. (2018). *Econometric analysis* (8th ed.). Pearson.

Hanandita, W., & Tampubolon, G. (2014). Does poverty reduce mental health? An instrumental variable analysis. *Social Science & Medicine*, *113*, 59–67. https://doi.org/10.1016/j.socscimed.2014.05.005

Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, *50*(2), 373–419. https://doi.org/10.3368/jhr.50.2.373

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.

Sullivan, A. L., & Field, S. (2013). Do preschool special education services make a difference in kindergarten reading and mathematics skills?: A propensity score weighting analysis. *Journal of School Psychology*, *51*(2), 243–260. https://doi.org/10.1016/j.jsp.2012.12.004

White, M. P., Alcock, I., Wheeler, B. W., & Depledge, M. H. (2013). Would you be happier living in a greener urban area? A fixed-effects analysis of panel data. *Psychological Science*, *24*(6), 920–928. https://doi.org/10.1177/0956797612464659