2015

A note on evaluating Supplemental Instruction

Alfredo R. Paloyo

University of Wollongong, apaloyo@uow.edu.au

A note on evaluating Supplemental Instruction

A note on evaluating Supplemental Instruction

Alfredo R. Paloyo

ABSTRACT
Selection bias pervades the evaluation of supplemental instruction (SI) in non-experimental settings. This brief note provides a formal framework to understand this issue. The objective is to contribute to the accumulation of credible evidence on the impact of SI.

INTRODUCTION
In a recent systematic review on the effectiveness of Peer-Assisted Study Sessions (PASS) or Supplementary Instruction (SI), Dawson, van der Meer, Skalicky, and Cowley (2014, p. 609) concluded that this kind of academic support is "correlated with higher mean grades, lower failure and withdrawal rates, and higher retention and graduation rates." Their conclusion is based on the collected body of evidence in recent years on SI. The purpose of this brief note is to provide a formal framework to demonstrate the problem of selection bias which pervades the majority of research on SI, including some of the published research that Dawson et al. (2014) cite.[1]

The intended audience is educational researchers who wish to conduct their own evaluation or to understand the weaknesses of existing evidence. The objective is to ultimately contribute to the effort to accumulate credible evidence on the impact of SI on a number of interesting outcomes, including not just final marks, but also perhaps non-traditional outcomes such as lecture attendance and student satisfaction.

THE EVALUATION PROBLEM
Naive impact evaluation typically involves a comparison of observed mean outcomes between those who received the treatment and those who did not. In the context of the present manuscript, for example, the average or mean final marks of SI participants and nonparticipants may be obtained, and the difference between the two is used as an estimate of the impact of SI. Unfortunately, this approach does not take into account the fact that participation in SI is typically a voluntary decision, and, as such, is influenced by individual characteristics—observed and (crucially) unobserved to the program evaluator—that may also contribute to the final mark.

The typical example is innate but unobserved motivation or ability, which may influence both the student's likelihood to participate in SI and her final

---

[1] An annotated bibliography on peer learning outcomes is available from http://z.umn.edu/peerbib.

mark. This confounds the estimate of the program impact (i.e., the exclusive impact of SI) obtained from a simple comparison of means. We refer to this confounding effect as the self-selection bias.

To formalise ideas, suppose one is interested in the impact of SI on final marks. The following discussion is primarily based on Angrist and Pischke (2009), but one may also refer to Holland (1986) for an earlier treatment from a statistics perspective. Let y_i denote the outcome for individual i, and let d_i denote a binary indicator variable that equals 1 if individual i received SI. Before the receipt of SI, an individual has two potential outcomes: $y_i(0)$ and $y_i(1)$, representing the potential outcomes without SI and with SI, respectively. However, after the delivery of SI, we only observe $y_i = y_i(d_i) = y_i(0)(1 - d_i) + y_i(1)(d_i)$. That is, only one of the potential outcomes can ever be realised. This is the well-known "fundamental problem of causal inference" (Holland, 1986) which arises since we are unable to observe the counterfactual situation for any single individual.

In terms of conditional expectations, one can show that

$$\begin{aligned}
\mathrm{E}[y_i|d_i = 1] &- \mathrm{E}[y_i|d_i = 0] \\
&= \{\mathrm{E}[y_i(1)|d_i = 1] - \mathrm{E}[y_i(0)|d_i = 1]\} \\
&+ \{\mathrm{E}[y_i(0)|d_i = 1] - \mathrm{E}[y_i(0)|d_i = 0]\}.
\end{aligned} \tag{1}$$

As Angrist and Pischke (2009) explain, the difference in outcomes between those who received SI and those who did not (receive SI) consists of two components. First, there is the impact of SI on those who actually received SI. This is the first pair of terms inside braces (note the conditioning on $d_i = 1$), and this is usually called the average treatment effect on the treated. Second, there is the selection-bias term, which is the pair contained in the second braces.

In the context of evaluating the impact of SI on student outcomes, we expect the selection-bias term to be nonzero, implying that the observed difference in, say, final marks is not equal to the effect of SI because it is contaminated by self-selection. Good final marks can be expected from motivated students, but motivation is also positively correlated with participation in SI. This implies that the following inequality holds: $\mathrm{E}[y_i(0)|d_i = 1] > \mathrm{E}[y_i(0)|d_i = 0]$; that is, the bias term is positive. In other words, without taking selection into account, one would overestimate the impact of SI using a basic comparison of mean outcomes between participants and non-participants.[2]

One way to ensure that the selection bias is actually zero is to randomise the provision of SI to the students. In that case, the potential final marks would be independent of treatment status. By design, the researcher can eliminate

---

[2] That motivation should be controlled for is highlighted in a number of previous studies (e.g., Gattis, 2002).

the selection bias in Equation (1) by random assignment. This means that participation in SI is no longer correlated with individual observed and unobserved characteristics. This explains why randomised controlled trials still constitute the "gold standard" for impact evaluation.[3]

However, controlled trials are difficult to implement, especially outside the clinical or laboratory setting. Unlike bacteria in a petri dish, there are major ethical and practical considerations in social experiments. Consider, for instance, the fact that there is some evidence that SI can improve student outcomes. It would be difficult to ethically justify depriving a random group of students access to SI simply because we want to evaluate its impact.

Nonetheless, there are a number of quasi-experimental approaches that still provide credible impact estimates under certain conditions. Examples of quasi-experimental approaches are instrumental-variables estimation, difference-in-differences, and regression-discontinuity designs. In the context of evaluating SI, these methods are particularly useful, especially since a randomised experiment may not be possible because of ethical or practical reasons.[4]

CONCLUSION

The evaluation of PASS or SI based on experimentally-generated data is rare. The majority of the literature on the topic relies on evidence obtained from non-experimental approaches that fail to account for the presence of self-selection bias. This note discusses how this bias causes problems in impact evaluation. The hope is that education researchers, especially those who are interested in estimating the impact of SI, can use this note to justify the use of experimental or quasi-experimental methods and to enable them to be critical of weak evidence. Ultimately, this will enable education researchers to contribute to a larger body of credible evidence on the impact of SI on a number of interesting outcomes.

REFERENCES

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Dawson, P., van der Meer, J., Skalicky, J., & Cowley, K. (2014). On the effectiveness of supplemental instruction: A systematic review of Supplemental Instruction and Peer-Assisted Study Sessions literature between 2001 and 2010. *Review of Educational Research, 84*(4), 609–639.

Gattis, K. W. (2002). Responding to self-selection bias in assessments of academic support programs: A motivational control study of Supplemental Instruction. *Learning Assistance Review 7*(2), 26–36.

---

[3] In this setting, one does not formally need additional control variables. However, their inclusion could greatly reduce the sampling error, resulting in more precise estimates (i.e., lower standard errors). Naturally, increasing the sample size will also increase the precision of the estimates.

[4] The most important non-experimental evaluation techniques are discussed in, for example, Gertler et al. (2011), Imbens and Wooldridge (2009), TenHave et al. (2003), and West et al. (2008). A detailed discussion of these methods is beyond the scope of this brief note. We refer the interested reader to the references provided.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2011). *Impact evaluation in practice.* Washington, DC: The International Bank for Reconstruction and Development/The World Bank.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5–86.

TenHave, T. R., Coyne, J., Salzer, M., & Katz, I. (2003). Research to improve the quality of care for depression: Alternatives to the simple randomized clinical trial. *General Hospital Psychiatry, 25*(2), 115–123.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Dolan Mullen, P. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health, 98*(8), 1359–1366.