Evaluation and Program Planning 47 (2014) 26–34

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Sensitivity analyses for clustered data: An illustration from a large-scale clustered randomized controlled trial in education

Yasuyo Abe [a,1], Kevin A. Gee [b,*]

[a] IMPAQ International, 1333 Broadway, Suite 300 , Oakland, CA 94612, United States
[b] University of California, Davis, One Shields Ave., Davis, CA 95616, United States

ARTICLE INFO

Article history:
Received 13 November 2013
Received in revised form 30 June 2014
Accepted 7 July 2014
Available online 12 July 2014

Keywords:
Impact evaluation
Sensitivity analyses
Data analysis for clustered randomized trials
Hierarchical linear modeling
Generalized estimating equations

ABSTRACT

In this paper, we demonstrate the importance of conducting well-thought-out sensitivity analyses for handling clustered data (data in which individuals are grouped into higher order units, such as students in schools) that arise from cluster randomized controlled trials (RCTs). This is particularly relevant given the rise in rigorous impact evaluations that use cluster randomized designs across various fields including education, public health and social welfare. Using data from a recently completed cluster RCT of a school-based teacher professional development program, we demonstrate our use of four commonly applied methods for analyzing clustered data. These methods include: (1) hierarchical linear modeling (HLM); (2) feasible generalized least squares (FGLS); (3) generalized estimating equations (GEE); and (4) ordinary least squares (OLS) regression with cluster-robust (Huber–White) standard errors. We compare our findings across each method, showing how inconsistent results – in terms of both effect sizes and statistical significance – emerged across each method and our analytic approach to resolving such inconsistencies.

1. Introduction

Cluster randomized controlled trials[1] (RCTs) have become an increasingly popular way to evaluate the impact of interventions which are applicable to intact groups of individuals. Common examples include schools that are randomly assigned to offer its students an educational intervention. Similarly, there are studies in which clinics are randomized to offer a particular treatment to an intact group of patients it serves. One notable feature of such trials is that individuals (e.g., students or patients) are clustered together in higher level units (e.g., schools or clinics) with the higher level unit serving as the unit of randomization[2]. Evaluators who analyze data from clustered RCTs must select from a variety of methods that appropriately account for the correlation between study participants within the higher level units. Ignoring such correlation, especially when the correlation between individuals within clusters is relatively high (as captured by the intra-class correlation coefficient (ICC)) may lead to erroneous inferences due to downward biased standard errors (Garson, 2012; Hox, 2010; Liang & Zeger, 1993; Zyzanski, Flocke, & Dickinson, 2004).

For evaluation analysts, deciding upon which method to use when analyzing clustered data is not an exact science. Often, the choice depends upon a combination of factors including analysts' professional judgment and their prior quantitative training. Also, the choice of method is driven by the methodological conventions and traditions of the disciplinary field (e.g., public health, education, etc.) in which the evaluation is conducted. However, one overarching principal is that analysts are entrusted to choose the most appropriate approach among various data analytic methods, *prior to conducting analyses*, based on their prior assessment of the design and data limitations. This prevents researchers from selecting, or being suspected of selecting, a particular analytic method to influence the results.

Yet, when accounting for clustering, analysts often rely only upon one preferred methodological approach without considering how and if the results remain consistent across different methods. Carrying out analyses using different methods and checking for the consistency in results across such methods is one class of a broader set of *sensitivity analyses* (Thabane et al., 2013) which analysts often undertake. We believe that well-thought-out sensitivity

* Corresponding author. Tel.: +1 530 752 9334.
E-mail addresses: yabe@impaqint.com (Y. Abe), kagee@ucdavis.edu (K.A. Gee).
[1] Tel.: +1 510 465 7884.

[1] *Cluster* randomized trials are also commonly referred to as *place-based* or *group* randomized trials (Boruch, 2005, p. 14).

[2] Often, these clustered structures (e.g., students within schools) are also referred to as multi-level, hierarchical or nested structures.

http://dx.doi.org/10.1016/j.evalprogplan.2014.07.001
0149-7189/

analyses to handle clustered data and the transparent reporting of such analyses are important, particularly as different methods can and – as we show in our case – lead to discrepant findings. When conflicting findings emerge across different methodological approaches, we believe that evaluation analysts must then proceed to understand the conflicting results, plan alternate analyses to reconcile such findings, and carefully document those alternative approaches. Finally, analysts should be transparent in communicating their analytic decisions to their evaluation audience.

In this paper, we review our results from a recently completed cluster randomized trial of a teacher professional development program. We compare our results across four methods we used to account for clustering in our data: (1) hierarchical linear modeling (HLM); (2) feasible generalized least squares (FGLS); (3) generalized estimating equations (GEE); and (4) ordinary least squares (OLS) regression with robust clustered (Huber–White) standard errors. Importantly, we show how inconsistent results emerged across these different methods and our approach to resolving inconsistencies. We present and discuss our work primarily from an *applied* point of view, forgoing technical descriptions of the methods we have employed (with the exception of the statistical model we present for our main analytic approach using HLM). We do assume, however, that readers have basic familiarity with statistical concepts and the analytic issues that arise due to clustered data.

We structure the rest of our paper in five sections. In Section 2, we briefly review clustered randomized controlled trials and introduce the concept of the *intra-cluster correlation coefficient* (ICC). The ICC is a key quantitative measure capturing the extent to which individuals are correlated within an intact group. We also discuss sensitivity analyses for clustered data, methods for handling clustered data and prior empirical studies that have compared methods for clustered data. Next, in Section 3, we describe our research design, providing background about our study intervention, the site and sample as well as our data and measures. In Section 4, we describe our primary analytic method along with our selected alternative methods. Then, in Section 5, we present results from the four analytic approaches we used to analyze our data, discussing the inconsistencies that emerged across the methods and ways in which we reconciled those inconsistent results. Finally, in Section 6, we close with several substantive "lessoned learned" of our work, providing advice to evaluation analysts who face the task of analyzing clustered data.

2. Clustered randomized trials and clustered data

A *cluster* randomized controlled trial (RCT) refers to an experiment in which intact groups of individuals are randomly assigned to receive an offer to participate in a treatment or not[3]. The groups that do not receive an offer of the treatment serve as the control group. This is in contrast to a standard RCT in which *individuals* are randomly assigned into a treatment or control group. The level at which randomization occurs – whether it be at the group or the individual-level – is commonly referred to as *the unit of randomization*. The cluster is the unit of randomization in numerous experimental evaluations of programs in education, public health and criminology (Boruch, 2005). Randomizing clusters of individuals not only avoids potential cross contamination between control and treatment conditions, but the interventions themselves are often designed to be administered to intact groups rather than individuals (Raudenbush, 1997). Finally, there may be ethical issues that can be ameliorated by randomizing at the cluster level. For example, in a ground-breaking study of an incentive-based cash subsidy program in Mexico known as *Progresa* (now *Opportunidades*), intact communities rather than households were randomly assigned to receive an offer of a subsidy or not (Parker & Teruel, 2005). Randomizing households within these relatively small and close-knit communities could have created tension between treatment and control group households (Parker & Teruel, 2005). Also, randomization could have led to a "perception of discretionality" (Parker & Teruel, 2005, p. 208) with respect to which households – despite being equally eligible – were selected to receive subsidies or not.

When analyzing data from cluster RCTs, evaluation analysts often want to understand the impact of a program, on average, *across individuals'* outcomes even though these individuals are part of an existing intact group. For example, in the evaluation of the *Progresa* program, researchers wanted to understand whether children living in communities randomized to receive cash subsidies had improved health outcomes versus children in control communities (Gertler, 2004). To determine the impact of the program on individuals' outcomes in a RCT with individual-level random assignment, an analyst may apply standard *t*-tests to compare the means of outcome measures collected on individuals assigned into the control condition versus the treatment condition or to apply ordinal least squares (OLS) regression to test the estimate effects on the treatment condition. However, such a strategy, if applied to a cluster RCT, ignores the fact that individuals are members of existing groups and may not be completely independent of each other—a critical assumption of standard statistical techniques such as the *t*-test or OLS regression. Ignoring clustering can lead to erroneous inferences (Garson, 2012; Hox, 2010; Liang & Zeger, 1993; Zyzanski et al., 2004) due to standard errors that are biased downwards (Clarke, 2008; Steenbergen & Jones, 2002) leading to inflated Type I error rates (i.e., stating that there is an effect when there is not). Modeling the degree to which individuals are correlated within clusters requires different methods, such as the ones we illustrate in this paper.

The degree to which individuals are interdependent within a cluster can be quantitatively measured by the intra-class correlation coefficient (ICC), often denoted by the Greek symbol ρ (rho) (Killip, Mahfoud, & Pearce, 2004). In the most basic case where individuals (e.g., students) are clustered into higher level units (e.g. schools), the ICC is calculated as the ratio of the between-cluster variance on a particular continuous outcome measure of interest (e.g. achievement) to the total variance (the between-plus within-cluster variance) of that outcome. The ICC can be expressed as:

$$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}} \quad (1)$$

where $\sigma^2_{\text{between}}$ represents the between-cluster variance and σ^2_{within} is the within-cluster variance. The ICC ranges from 0 to 1, with values closer to 1 indicating a higher degree of correlation for a particular outcome of interest within an intact group[4]. If there is no variability between clusters, the ICC would equal 0 $(0/(0+\sigma^2_{\text{within}}) = 0)$. This suggests that individuals' outcomes are independent of each other. In other words, all of the variation lies between individuals and there is no correlation between individuals within a cluster. On the other hand, in the instance where all individuals are homogenous on an outcome so that there is no within-cluster variance, the ICC would equal 1 $(\sigma^2_{between}/(\sigma^2_{between} + 0) = 1)$. The ICC can also be interpreted in

[3] Here, we assume the most basic design of a randomized experiment with only one treatment and one control condition. There are, of course, various randomized designs that have multiple treatment and control conditions.

[4] There are instances in which the ICC can be negative (see Lohr, 2010, p. 175); however, as Lohr (2010) notes, "The ICC is rarely negative in naturally occurring clusters" (p. 175).