Gifted Child Quarterly
54(2) 152–155

DOI: 10.1177/0016986210363076
http://gcq.sagepub.com
SAGE

Dealing With Dependence (Part I): Understanding the Effects of Clustered Data

D. Betsy McCoach[1] and Jill L. Adelson[2]

Abstract

This article provides a conceptual introduction to the issues surrounding the analysis of clustered (nested) data. We define the intraclass correlation coefficient (ICC) and the design effect, and we explain their effect on the standard error. When the ICC is greater than 0, then the design effect is greater than 1. In such a scenario, the standard error produced under the assumption of independence is underestimated. This increases the Type I error rate. We provide a short illustration of the effect of non-independence on the standard error. We show that after accounting for the design effect, our decision about the statistical significance of the test statistic changes. When we fail to account for the clustered nature of the data, we conclude that the difference between the two groups is statistically significant. However, once we adjust the standard error for the design effect, the difference is no longer statistically significant.

Keywords

interclass correlation, clustered data, hierarchical linear modeling, gifted education research

In this Methodological Brief, Dr. Betsy McCoach (University of Connecticut) and Dr. Jill Adelson (University of Louisville) provide an overview of the use of hierarchical linear modeling with nested data. This is the first piece in a two-part series that focuses on HLM and we hope readers find it useful as they work with nested data.

Tonya R. Moon, PhD
GCQ Methodological Briefs, Editor
Associate Editor, *GCQ*

In traditional statistics courses, we learn a variety of inferential statistical techniques, each of which makes the assumption of independence of observations. The assumption of independence means that cases "are not paired, dependent, correlated, or associated in any way" (Glass & Hopkins, 1996, p. 295). However, in educational research, we rarely collect data that meet this stringent assumption. For instance, students who receive instruction together in the same classroom, delivered by the same teacher, tend to be more similar in their achievement (and other educational outcomes) than students who were instructed by different teachers. Generally speaking, observations that are clustered tend to exhibit some degree of interdependence. Examples of clustering include students clustered within classes, teachers clustered within schools, children clustered within families, or even observations across time, which we consider to be clustered within persons.

Having clustered data presents researchers with several problems, both conceptual and statistical. Conceptually, the researcher may be interested in studying relationships among variables that occur at multiple levels of the data hierarchy as well as potential interactions among them. For instance, a researcher may wish to study the relationships between student ability, teaching style, and academic achievement. Student ability is measured at the individual level, whereas teaching style is measured at the classroom/teacher level. There is a potential moderating effect of the teacher's teaching style on the impact of student ability on student achievement. For example, behaviorist teachers may be more effective with low-ability students and less effective with high-ability students. In contrast, constructivist teachers may be more effective with high-ability students and less effective with low-ability students. Therefore, the relationship between ability and achievement would be stronger in constructivist classrooms than it would be in behaviorist classrooms. These are the types of questions that multilevel analyses are able to address. Using traditional methodologies such as regression

[1]University of Connecticut, Storrs, CT, USA
[2]University of Louisville, Louisville, KY, USA

Corresponding Author:
D. Betsy McCoach, Educational Psychology Department, University of Connecticut, 249 Glenbrook Road, Unit 2064, Storrs, CT 06269-2064, USA
Email: betsy.mccoach@uconn.edu

and ANOVA does not allow the researcher to ask and answer same sort of nuanced research questions that span across the different levels.

In addition, nested data present researchers with statistical challenges. Although researchers traditionally make the assumption that their observations are independent, having nested data violates this assumption. Research consistently has demonstrated that students in one cluster (such as a class, school, etc.) tend to be more similar to each other in terms of an outcome variable (such as achievement) than they are to students in another cluster. This interdependence, which is a result of the sampling design (choosing to study students within classes or students within schools, for example, rather than taking a random sample of students without regard to the classroom or school in which they are enrolled), affects the variance of the outcome, which in turn affects the estimates of the standard errors. Ignoring this clustering effect or nonindependence, as we often do in traditional tests of significance, incorrectly reduces the standard error, artificially increases the confidence in our estimates or associations with the outcome, and, thus, increases the possibility of making a Type I error (rejecting the null hypothesis when we should have failed to reject it) (O'Connell and McCoach, 2008).

Fortunately, researchers are able to address these conceptual and statistical issues using multilevel modeling, also commonly referred to as hierarchical linear modeling, or HLM (Raudenbush & Bryk, 2002). HLM has many benefits. The standard errors from HLM analyses take into account the clustered nature of the data, resulting in a more correct Type I error rate. In addition, HLM allows researchers to model multiple levels of a hierarchy simultaneously, partition variance across the levels of analysis, and examine relationships and interactions among variables that occur at multiple levels of a hierarchy. The remainder of this article addresses the statistical issue of nonindependence in clustered data and the implications this has on traditional tests of significance, illustrating the benefit of using HLM to model the nested nature of much of educational research.

Treating clustered data as if they were independent data results in standard errors that are underestimated. Conceptually, this is because when people exhibit some level of homogeneity, there is some redundancy (or correlation) in their responses, and that redundancy can be explained or accounted for by their cluster membership. Because of this redundancy, the "effective sample size" for the study is smaller than the actual sample size for the study. The degree to which the effective sample size and the actual sample size differ affect the degree to which standard errors from traditional statistical tests are underestimated. To get a rough estimate of the degree to which the clustered nature of the data affects the standard errors, it is necessary to take into account the degree of homogeneity within clusters as well as the average cluster size.

The Intraclass Correlation

The *intraclass correlation (ICC)* provides a measure of how similar, or homogeneous, individuals are within clusters. The ICC is the proportion of variability in the outcome that is accounted for by the clusters or groups. To calculate the ICC, often referred to as ρ (rho or roh), we partition the total variability into two pieces: that which is within clusters (σ^2) and that which lies between clusters (τ_{00}). To compute the ICC, we simply divide the between-cluster variability (τ_{00}) by the total variability ($\tau_{00} + \sigma^2$), as the following formula shows:

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2}.$$

A large ICC indicates that there is a large degree of homogeneity within clusters (σ^2 is small) and/or a large degree of heterogeneity across clusters (τ_{00} is large). A recent review by Hedges and Hedberg (2007) indicates that when the school represents the cluster variable, the average ICC for student academic achievement (in either mathematics or reading) is .22. After controlling for pretest scores and/or demographic characteristics such as socioeconomic status and so on, the average ICC is .11 to .12.

Using the ICC, we calculate a *design effect* (*DEFF*; Kish, 1965). The *DEFF* is a ratio of the sampling variability for the study design compared with the sampling variability that would be expected if the study used a simple random sample (SRS) and can be calculated using the following equation:

$$\frac{\text{var}(design)}{\text{var}(SRS)} = 1 + \rho(\bar{n}_j - 1),$$

where n_j is the average sample size within each cluster and ρ is the ICC. If this ratio is equal to 1, then there is no clustering effect. However, if it is greater than 1, the research design has violated the assumption of independence of observations, which would lead to bias in traditional tests of significance. The design effect can then be used to calculate the *effective sample size*, or the sample size that we should use to estimate the standard error for the study. The *N* effective is simply (Snijders & Bosker, 1999) the following:

$$\frac{N}{DEFF} = \frac{N}{1 + \rho(\bar{n}_j - 1)}.$$

To address this violation, the researcher could use HLM or could use the *DEFT*, the square root of the *DEFF*, to adjust the standard errors. The *DEFT* indicates the degree to which the standard errors need to increase to account for the

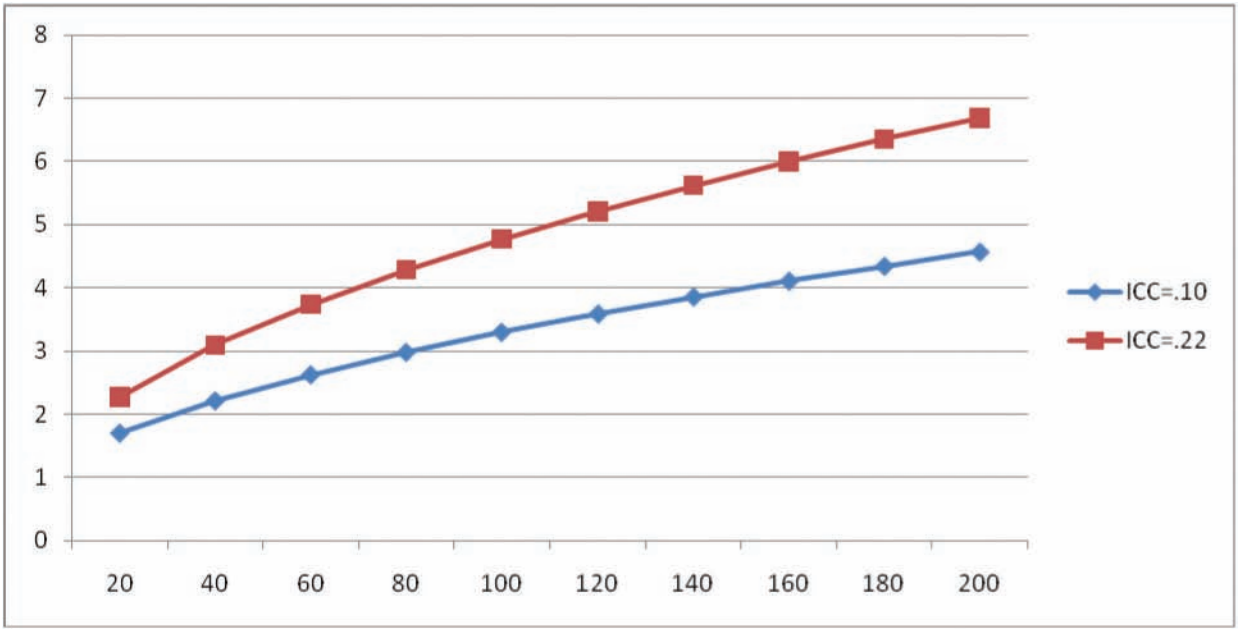

Figure 1. The DEFT as a function of n_j, with the ICC held constant at .10 or .22.

clustering.[1] Generally, *DEFT* increases both as the average cluster size increases and as the ICC increases. Often, in school-based research, the cluster is the classroom. Even with an ICC as small as .15 (see Figure 1) and average cluster size of 25, the standard errors have significant bias. In this case, $DEFT = \sqrt{1 + .15 \cdot (25 - 1)} = 2.14,$ which indicates that the standard errors assuming a simple random sample are less than half as large as they should be if we took the clustering into account. The effect is even more pronounced when the school is the cluster because the average sample size at the school level is often quite large. As mentioned before, the typical ICC for school effects research on achievement is .22. With an ICC of .22 and an average cluster size of 100, $DEFT = \sqrt{1 + .22 \cdot (100 - 1)} = 4.77,$ indicating that the standard errors produced by standard software programs are close to five times smaller than they should be if they took the clustered nature of the data into account.

Illustration of the Design Effect With Nested Gifted Education Data

To illustrate the impact of the ICC on statistical significance, as well as how to demonstrate how to use the *DEFT* to correct standard errors for the degree of clustering, we use actual Peabody Picture Vocabulary Test data from an intervention study (Coyne, McCoach, Zipoli, & Ruby, in press).

In this small data set, 121 students were nested within eight classrooms. Seventy-eight students received a vocabulary intervention, whereas 43 served as control students. There were not enough clusters in the data set to run a multilevel analysis; however, the researchers wished to account for the nonindependence in their data set when they were computing the statistical tests. Therefore, they began by running an unconditional random-effects ANOVA model to partition the total variance into between-classroom variance (13.25) and within-classroom variance (168.98). They then divided the between-classroom variance (13.25) by the total variance (168.98 + 13.25) to compute the ICC (.078). Next, they computed the *DEFF*, which is $1 + \rho(\bar{n}_j - 1)$. In this case, 121 students were nested within eight classes, which made the average cluster size 15.125. Therefore, the *DEFF* = 1 + .078(15.125 − 1) or 2.10. Next, to obtain the *DEFT*, they took the square root of the *DEFF*, which is 1.45. A *DEFT* of 1.45 means that they should multiply the standard errors from other traditional statistical analysis by a factor of 1.45 to get a more realistic estimate of the standard errors, given the dependence in the data.

Next, the researchers ran a *t* test comparing the treatment and control groups on the Peabody Picture Vocabulary Test. The researchers' naive analysis of the difference between the treatment and control groups was statistically significant. The mean difference between the treatment and control groups was 6.48; the standard error was 2.40 (t = 2.70, p = .008). Next, they applied the correction for the DEFT to the standard error and recomputed the *t* ratio and *p* value. The corrected standard error was 2.40 × 1.45 = 3.48. This resulted in a corrected *t* ratio of 6.48/3.48 = 1.86, corresponding to a *p*

value of .07. Therefore, the results were no longer statistically significant at $p = .05$. This example illustrates the impact that even a modest ICC can have. Given that ICCs in teacher and school effects research often lie within the .10 to .25 range, it is easy to see how ICCs of this magnitude could change the inferences that we make about the statistical significance of a given study.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

Note

1. It is worth noting that using the *DEFT* to correct standard errors is an approximation technique and that more exact methods of computing standard error estimates such as using Taylor series expansion, balanced repeated replications, and bootstrap methods, as are implemented in specialized statistical software packages such as SUDAAN, WESVAR, SPSS Complex Samples, and AM, provide more accurate estimates of standard errors.

References

Coyne, M., McCoach, D. B., Zipoli, R., & Ruby, M. (in press). Direct and Extended Vocabulary Instruction in Kindergarten: Investigating Transfer Effects. *Journal of Research on Educational Effectiveness.*

Glass, G. V., & Hopkins, K. D. (1996). *Statistical Methods in Education and Psychology.* Boston: Allyn & Bacon.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60-87.

Kish, L. (1965). *Survey sampling.* New York: John Wiley.

O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel modeling of educational data.* Charlotte, NC: Information Age.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Snijders, T. A., & Bosker, R. J. (1999). *Multilevel analysis.* Thousand Oaks, CA: Sage.

Bios

D. Betsy McCoach is an associate professor in the Measurement, Evaluation and Assessment program at the University of Connecticut, where she teaches coursework in hierarchical linear modeling, structural equation modeling, instrument design, and research design. Her research interests include the underachievement of academically able students, growth curve modeling, and model fit issues. She has coedited the book *Multilevel Modeling of Educational Data,* and she is currently authoring a textbook on instrument design. She has published numerous peer review journal articles and book chapters in the areas of gifted education, research methodology, and educational research and currently serves as the research methodologist for several federally funded projects. She is the current coeditor of the *Journal of Advanced Academics*. She also serves on the editorial review boards for the *American Educational Research Journal*, the *Journal of Educational Psychology*, the *Journal of Educational Research*, and *Gifted Child Quarterly*. She is the 2007 recipient of the National Association for Gifted Children (NAGC) Early Scholar award and is currently the chair elect of NAGC's Research and Evaluation Network.

Jill L. Adelson, PhD, is an assistant professor in the Department of Educational and Counseling Psychology at the University of Louisville, where she teaches courses in educational statistics and measurement. Jill earned her doctorate in measurement, evaluation, and assessment and in gifted education from the University of Connecticut. Her research includes applying advanced methodologies, such as hierarchical linear modeling, propensity score analysis, and structural equation modeling. Her substantive interests include the effects of gifted programming, the talent development of mathematically talented elementary students, attitudes towards mathematics, and special issues for mathematically talented females. She serves in leadership positions in both the National Association for Gifted Children and the American Educational Research Association.