Losing Control: Conducting Studies with Comparison Groups

Jan Norton
University of Wisconson
Oshkosh

Studies in education often report the differences between participants' and non-participants' test scores, course grades, retention, and other criteria. When participants' average performance is higher, it can be difficult to attribute the improvements to participation. Comparing participants and non-participants on other measures can strengthen the argument that participation had a positive impact when the two groups are otherwise similar in relevant ways. Reviewing students' demographic characteristics, incoming ACT/SAT scores, previous grades, and placement results can establish points of comparison; specific statistics can also assist in identifying similarities and differences between groups.

In a research-perfect world, there would be control groups. We would review the pass rates in College Algebra according to whether students took a developmental math course or were denied such preparation. We would evaluate the effectiveness of tutoring by comparing grades of tutored students with those who were not allowed to receive assistance. We would randomly select participants for Supplemental Instruction so we could be more certain that SI attendance, not student motivation, had an impact on higher grades.

However, for most educators, such denial of services is ethically unacceptable. We believe that services like the above have a positive impact on student success, even if we can't prove and replicate results with the supposed predictability of traditional control-group research design. But an alternative is possible – and increasingly necessary – in education. Although we may not randomly select a group of students to participate in a study and a control group of students who do not participate, we can still compare the two groups in order to examine the benefits of participation.

A comparison group is simply a group of non-participating students who are similar in one or more important ways to the group of students being studied. Unlike a traditional experimental control group which is established prior to a study, comparison groups can be determined after the participants have self-selected by attending tutorial sessions, taking a developmental course, or becoming involved in the program or activity being evaluated. Assessing and reporting relevant similarities between participants and non-participants strengthens an argument that differences in other measures (e.g., improved grades or retention rates) may be due to participation.

Before looking at comparing students, it is worth noting that even in traditional control-group research, members of the control group – those who do not participate or receive the "treatment" – should be examined to determine whether they are comparable to those who did participate (Isaac & Michael, 1995). While random selection usually assures that key characteristics in both groups are equivalent, sometimes they are not. Looking at characteristics to determine group similarity is a legitimate part of traditional research methodology.

There are several ways to compare groups of students, and the characteristics you select depend on both the information available and the focus of your study. Many similarities can be examined using student demographics such as age and gender. You can also compare students using measures such as their ACT/SAT scores or their high school ranks. Comparisons based on grades in a course being studied can also reveal similarities between participants and non-participants. Most of these comparisons can be established using simple averages, but when the percentiles or means look dissimilar, some calculations such as t-tests and analysis of covariance can usually determine whether differences between groups are statistically significant.

Comparison Groups Based on Demographics

One way to compare groups of students is to examine one or more demographic characteristics. In general, demographics provide a measure of normalcy: you can use characteristics such as age, ethnicity, residency, gender, etc. to determine whether your participants and non-participants represent the student population overall. For example, if 30 percent of the participants are male, you could check to see if approximately 30 percent of the non-participants are male. Are the two groups comparable in age and/or socioeconomic status? Reviewing the demographic characteristics of successful participants also helps you discover whether different groups of students benefit equally from tutoring or developmental coursework or other assistance models.

There are numerous student characteristics that may be worth examining for the program you want to evaluate, as long as the demographic characteristics selected are meaningful. That is, the points of comparison must be relevant to the program being assessed (Fraenkel & Wallen, 2003) as well as the campus as a whole. For example, if returning adult students on your campus tend to earn better grades and you discover that participants in your math tutoring program earn better grades in College Algebra, you would want to examine the ages of your participants and non-participants in order to determine whether the two groups are comparable.

Comparison Groups Based on Incoming Measures

Most campuses collect academic information on students during the admissions process. If your institution requires ACT or SAT scores, each of these measures yields one primary score and several test scores that may be relevant. ACT scores include an overall composite and four test scores (English, Math, Reading, Science); the SAT Reasoning Test yields a total score that combines three test scores (Critical Reading, Math, Writing). For students without ACT/SAT scores, placement tests can also establish an incoming measure of a student's academic preparedness.

With any measurement instrument, select the most relevant score or subscore for your study. In the previous example about College Algebra in which you would review ages of participants and non-participants, ACT or SAT math scores would provide another basis for comparison. If you think that a voluntary common reading for students in a developmental reading course will result in better course grades, you might choose to use a reading test score as a point of comparison between the students who chose and those who declined the common reading experience. If you want to examine the impact of a service learning experience upon the students' evaluations of their freshman seminar, you would probably choose a more general composite or total score.

Keep in mind that the point of examining the scores is to establish whether the students who received the services and those who didn't are comparable in this specific way. Once you have selected the most relevant score to work with, calculate the mean scores for both participants and non-participants. If the means match or nearly match, you can reasonably assume that your results for participants and non-participants are not based on the students' existing abilities (as measured by the test). That is, you are reporting results for two groups of students who are, at least by this measure, equivalent in their overall academic abilities prior to the program or assistance you provided.

For example, in a Supplemental Instruction (SI) program that supports a chemistry course, grade differences are often the only data shared about the students in the course, as shown in Table 1.

Table 1

Course and ACT Scores by SI Participation

	Mean Chem grade	Mean ACT Science score
SI participants	3.15	21.2
Non-participants	2.63	20.9

Table 1 shows the standard ½ to 1 letter grade higher that is typically expected for SI participants. A frequent explanation for the grade differences argues the following: motivated students tend to perform better in classes; SI participation demonstrates motivation; therefore those students did better in the course. But adding the ACT science score to the overall picture shows that the level of preparation and motivation to learn science prior to the course was roughly equivalent for both groups. The ACT score similarity does not rule out the impact of motivation on grades, but it does minimize it somewhat. Additional comparisons could serve to further reinforce the similarity of the two groups.

If your institution collects students' high school ranking for course placement or program eligibility, that measure could be used as a relevant point of comparison. However, it is a relative number: students' abilities at one school may place them at the middle of their class, but those students' skills might qualify them as valedictorians in a less competitive environment. To a slightly lesser degree, high school GPAs are also relative and may vary widely in their ability to predict student preparedness for college. High school information is also less relevant for returning adult students.

Other incoming measures can contribute a more qualitative comparison of student characteristics. Noel Levitz's extensive *College Student Inventory* (CSI) provides a wealth of information on incoming freshmen, from measures of dropout proneness and predicted academic difficulty to a series of motivational scores including study habits, desire to finish college, and receptivity to support services (Stratil, 2004). The *Study Behaviors Inventory* (SBI) measures eight areas (e.g., time management, notetaking, writing, and faculty relations) to yield a score on three factors: academic confidence, short-term study behaviors, and long-term study behaviors (Bliss, Kerstiens, & Marvin, 1995). The *Learning and Study Strategies Inventory* (LASSI) yields a set of ten scores relevant to students' personal characteristics and their perceptions of their skills including motivation, time management, notetaking, and stress management (Weinstein, Palmer, & Schutte, 1987). Because an important difference between randomly selected control groups and self-selected groups is the motivation to participate in a treatment (Campbell & Stanley, 1963; Isaac & Michael, 1995), a motivation score from the LASSI or CSI could be a powerful point of comparison. While these three inventories are typically given during students' freshman year, the LASSI and SBI could be administered as part of an evaluation or research study.

Comparison groups based on course grades

If you are seeking to determine whether an educational intervention (tutoring, change in course delivery, SI, workshops, etc.) had an impact, you could look at grades in a previous course, if there is a clear relationship between the content of the two courses. For example, if the students in Math 2 also took Math 1, it is fairly reasonable to assume that their Math 1 grades provide some insight into their level of performance expected for Math 2, the next course in the sequence. In Table 2, both groups of students earned higher grades in Math 2 than they earned in Math 1; those who participated in tutoring received noticeably higher Math 2 grades.

Table 2

Math Course Grades by Participation in Tutoring

	Mean Math 1 grade	Mean Math 2 grade
Students who participated in Math 2 tutoring	2.75	2.89
Students who did not participate in Math 2 tutoring	2.42	2.57

While it may be tempting to focus only on the Math 2 grades, it looks as if the tutoring participants in Math 2 were going to do better in the class anyway: after all, their grades in Math 1 were higher. Did tutoring help? Yes, probably, but because the participants and non-participants weren't equal to begin with, it's harder to argue for the positive impact of tutoring.

However, in Table 3, when the Math 1 grades are comparable, the difference in the Math 2 grades appears to be more dramatic. The figures don't prove that tutoring had a positive impact, but it's a more believable argument than the figures in Table 2 would support.

Table 3

Math Course Grades by Participation in Tutoring

	Mean Math 1 grade	Mean Math 2 grade
Students who participated in Math 2 tutoring	2.38	2.89
Students who did not participate in Math 2 tutoring	2.42	2.57

Similar comparison groups can be examined using grades within a course. For example, if students take their first quiz in Economics before the learning center offers a workshop on “Taking Notes in Economics,” the first quiz scores provide a baseline performance in the class. If after the workshop, those who attended score noticeably better on the second quiz, the first scores help support the claim that the workshop was genuinely helpful.

Going Beyond Means

In many comparisons between participants and non-participants in a study, there will be one or more measures that can demonstrate how comparable the groups of students are. When the results of participation clearly show a difference and the comparison measures are clearly similar, those measures help support the contention that the participants’ improved scores are due to the intervention they received. But sometimes the similarities or differences between the groups of students are not easily discernable. In these circumstances, there are several additional steps you can take: calculate a t-test, calculate an analysis of covariance, or replicate the evaluation or research study.

Sometimes mean scores will look so close that it appears as if there are no measurable performance differences between two groups of students. When that occurs, a statistic known as the t-test may be able to indicate whether there are statistically significant differences between two means. Another statistic that may be useful is the analysis of covariance (ANCOVA). By incorporating additional information and statistically equalizing some of the differences in those factors, ANCOVA may be able to demonstrate meaningful differences in results even when the baseline measures are not similar. There are specific circumstances in which a t-test or ANCOVA is appropriate, and it is beyond the scope of this article to examine those subtleties. It is nevertheless important to be aware that even if a measure such as mean test scores or course grades do not seem to indicate that an intervention had any kind of impact, a t-test or analysis of covariance may help to establish a statistically significant difference.

At times the best way to demonstrate that students benefited from participation is to repeat the study. Even a minimal improvement for participants is made more meaningful when it can be achieved multiple times with different groups of students.

Conclusion

Opportunities for traditional control group research are rare in education. There are significant ethical questions about withholding assistance or avoiding new instructional techniques in order to establish a control group, but there are also ethical concerns about providing services for which there is no assured benefit. For most classroom instructors and program administrators, a simple valuable compromise requires going beyond noting an improved performance by students who participate in a study. Establishing that participants and non-participants share key characteristics can substantiate the positive results achieved by participants.

References

Bliss, L., Kerstiens, G., & Marvin, R. (1995). *Study behaviors inventory.* [Inventory]. Torrance, CA: Andragogy Associates.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

Fraenkel, J. R., & Wallen, N. L. (2003). *How to design and evaluate research in education* (5th ed.). Boston: McGraw-Hill.

Isaac, S., & Michael, W. B. (1995). *Handbook in research and evaluation: For education and the behavioral sciences* (3rd ed.). San Diego: Educational and Industrial Testing Services.

Stratil, M. (2004). *College student inventory.* [Inventory]. Iowa City: Noel Levitz, Inc.

Weinstein, C. E., Palmer, D. R., & Schutte, A. C. (1987). *Learning and study strategies inventory.* [Inventory]. Clearwater, FL: H & H Publishing.

Jan Norton directs the Center for Academic Resources at the University of Wisconsin Oshkosh. She has a Master's in Educational Research and Psychology. As a consultant, she helps learning centers evaluate their programs. Jan reviews program certification applications for NADE and leads CRLA's Research & Evaluation SIG. She welcomes your comments and questions at nortonj@uwosh.edu.