Teaching of Psychology
1-7
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00986283211022390
journals.sagepub.com/home/top
SAGE

Improving the Assessment of Teaching Effectiveness With the Nonequivalent Dependent Variables Approach

Joshua J. Reynolds[1]

Abstract

Introduction: Assessing teaching effectiveness is relevant for improving one's teaching and for moving through the tenure process; however, the validity of assessment methods, such as Student Evaluations of Teaching (SET), have been heavily criticized. **Statement of the Problem:** Using a one–group pretest–posttest design and assessing learning over the semester has several advantages over SET; however, one drawback is in making conclusions about the cause of changes in the post-test. A change could be due to learning in the semester, maturation, history, or even a testing effect. **Literature Review:** To improve the inferential quality of teaching assessment, a nonequivalent dependent variable (DV) design is highly advantageous. A nonequivalent DV is an outcome that is *not* the target of the intervention yet responds to the same contextually relevant factors. **Teaching Implications:** By using a nonequivalent DV design, there might be an increase from the beginning of the semester to the end of the semester in the main DV, but no increase in the nonequivalent DV, which provides a stronger argument that the change in the main DV is due to a true learning effect. **Conclusion:** Using nonequivalent DV methodology improves inferential quality and is easily implemented.

Keywords

teaching effectiveness, assessment, nonequivalent dependent variables design

Assessing teaching effectiveness aids in improving the quality of teaching but it is also important for the tenure process. There are a variety of strategies for measuring teaching effectiveness (Berk, 2005). Student Evaluations of Teaching (SET) are the most common form of assessment, but many have argued that SET lack validity and are vulnerable to bias (Boring et al., 2016; Esarey & Valdes, 2020; Hoorens et al., 2020; Sinclair & Kunda, 2000; Uttl et al., 2017). Strategies like pre/post-tests, which fall under the umbrella term learning outcomes, are an important supplement to the often-required SET. A common design to use in learning outcomes is the one–group pretest–posttest. An assessment is administered to students at the beginning of the semester (pre) and at the end (post). While methodologically this is an improvement over the post-test only design, as scores before and after the course can be compared, there are still some serious disadvantages. To address these disadvantages and improve the inferential quality of teaching assessment, a nonequivalent dependent variable (DV) design can be used. This deceptively simple approach allows the instructor to better differentiate a true increase in student learning from changes in scores over time due to other causes. The purpose of this article is to 1) discuss some of the advantages and disadvantages of the one-group pre-test-post-test design, 2) to show that the nonequivalent DV approach is highly advantageous, 3) to demonstrate how this can be implemented easily, and 4) to increase awareness about this type of design.

Learning Outcomes

One way to assess teaching effectiveness is to measure how much students learned over the semester, such as comparing their knowledge on multiple choice or essay items before and after they have been taught (a one-group pre-test-post-test design). In addition to knowledge, skills can also be assessed. For simplicity, this discussion will focus on assessing knowledge via multiple choice items.

Measuring knowledge in a one-group pre-test-post-test design offers several advantages over SET. For example, pre/post-tests allow the instructor to better understand the content areas students struggle with the most. Students are also exposed to the main topics in a course at the start of the semester and the types of items that might be used in exams during the semester

[1] Psychology Department, University of Scranton, PA, USA

Corresponding Author:
Joshua J. Reynolds, Psychology Department, University of Scranton, 800 Linden St., Scranton, PA 18505, USA.
Email: joshua.reynolds@scranton.edu

(e.g., synthesis, application). Another advantage of pre/post-tests over SET is that there is unlikely to be a contaminating effect of grade leniency. One concern with SET is that instructors are tempted to inflate student grades in exchange for more positive student evaluations, which shifts the focus of the class toward consumer-oriented teaching and not learning (Olivares, 2003). This grade leniency contamination is absent in pre/post-tests.

Research indicates that while SET primarily measures student's reactions to the instructor and course, pre/post-tests primarily measure learning and therefore different evaluation criteria (Arthur et al., 2003). While SET and pre/post-tests can be correlated, pre/post-tests are more strongly related to student grades than SET (Arthur et al., 2003; Stark-Wroblewski et al., 2007). Given the advantages of pre/post-tests, they are a useful addition to a comprehensive assessment of teaching effectiveness.

However, there are some methodological issues with the standard pre/post-test approach (i.e., one-group pre-test-post-test design). For example, maturation, or naturally occurring changes over time, can be a confound (Shadish et al., 2002). Students might naturally develop better study habits, take classes more seriously as the semester unfolds, or improve test taking skills, which could influence scores on the post-test. History effects are also possible. Students might experience some stressful events at the beginning of the semester, which would lower scores on the pre-test and then the increase in scores on the post-test could be falsely attributed to learning. The opposite is also possible, with scores on the post-test being artificially low. Regression to the mean effects is yet another issue; depending on the degree of random error on the pre/post-test, a student's low score on the pre-test is likely to be followed by a relatively higher score on the post-test, which is closer to that person's true mean. More problematic in the case of assessing learning are testing and practice effects. Testing effects, as described by Shadish et al. (2002), involve the influence of a previous test on a subsequent test. Students might improve on the post-test because they have previously been exposed to those items on the pre-test. One way to reduce this bias, at least to some extent, is randomize the order of the test items and answers on both the pre-test and post-test. This is easily accomplished using survey tools like Qualtrics or learning management systems like Blackboard. Alternatively, different items could be used for the pre-test and the post-test; however, this can result in an instrumentation confound (Shadish et al., 2002) and it is unclear if a change in scores is due to learning or a change in the instrument. Lastly, while an increase on the post-test could be genuinely attributed to learning, the learning may not be a product of teaching effectiveness but some other cause.

In addition to methodological issues, there is a statistical issue with the most common approach of analyzing pre/post-test scores. The most common approach in psychology is to conduct a frequentist Null Hypothesis Significance Test (NHST), namely the dependent or paired samples *t* test. This is a reasonable approach for analyzing such data; however, NHST has several serious issues. A full discussion of this is beyond the scope of this article but see Dienes (2011); Ioannidis (2019); Kruschke and Liddell (2018); Szucs and Ioannidis (2017); and Wagenmakers et al. (2018). One issue is that the typical .05 criteria is a relatively weak threshold for deciding whether there is an effect. In a within-subjects design, it is also easier to reject the null hypothesis given the increase in power. Further, *p* values are often misinterpreted, such as interpreting *p* values as the likelihood that the results are due to chance (Goodman, 2008).

In the case where the post-test has a higher average than the pre-test at $p < .05$, an instructor might consider this good evidence that students learned the material. However, this is misleading (see e.g., Aczel et al., 2017). Setting, a more stringent alpha (e.g., .001) can reduce Type I errors, but the probability of making a Type II error increases. Presenting an effect size estimate, such as Cohen's d and confidence intervals on that estimate, is also important, as *p* values do not contain effect size information. Unfortunately, like *p* values, confidence intervals are often misinterpreted (Kruschke & Liddell, 2018).

An alternative approach would be to use Bayes Factor. Unlike *p* values, the likelihood of the data for both the null and alternative hypothesis is considered in Bayes Factor. Thus, one interpretation of Bayes Factor is the strength of evidence in favor of one hypothesis/model compared to another. Bayes Factors yield effect size information, are more intuitive to interpret, and do not yield a dichotomous decision. In this case, they would allow the instructor to interpret how much evidence there is that there was a larger score on the post-test. A Bayes Factor approach, or another Bayesian approach, is arguably preferential in analyzing such data; however, statistical strengths of Bayesian approaches or other modern data analytic strategies, do not solve the methodological limitations, of the one–group pretest–posttest design.

There have been several methodological variations suggested for the typical pre/post-test. For example, Stark-Wroblewski et al. (2007) used 30 items taken from five tests given during the semester as a pre-test. Then, scores on the pre-test were compared against students' scores on those items in actual exams. An advantage of this approach is that the post-test is embedded into class time. A disadvantage of this approach is that in addition to not addressing the methodological weaknesses of the typical one–group pretest–posttest design, there is an additional confound introduced: motivation. Scores on the pre-test do not affect their grades, but scores on the post-test embedded items do affect their grades. Therefore, an increase in performance on the post-test embedded items might be attributed to motivation.

To minimize the testing confound and instrumentation confound, Bartsch et al. (2008) introduce the idea of using a one–group pretest–posttest design with alternative forms. In this approach, there are two versions of the test: A and B. Group one receives version A as the pre-test and version B as the post-test, while group two receives version B as the pre-test and version A as the post-test. Ideally, there would be random assignment of students for equivalence of the groups.

Similarly, in making the two versions of the test, random assignment of items would minimize systematic differences between the two versions. While the one–group pretest–posttest design with alternative forms was advanced as a design for assessing teaching demonstrations, it could be used as an assessment over the entire semester. There are some additional complexities in this design, such as how to keep track of who took what version of the test first. In small classes, equivalence of groups through random assignment is more likely to fail. Lastly, the additional coordination that is involved may reduce the likelihood that instructors use the design, despite its advantages over the standard one–group pretest–posttest design.

The Nonequivalent Dependent Variables (DV) Approach

The nonequivalent DV approach has been applied in research since at least the 1970s (e.g., McSweeny, 1978; Robertson & Rossiter, 1976). However, this approach is not heavily used, potentially because it is mentioned in relatively few methodology texts (Coryn & Hobson, 2011; see Brough (2019), Trochim and Donnelly (2016), and Shadish et al. (2002) for some notable exceptions). Shadish et al. (2002) defines a nonequivalent DV as "a dependent variable that is predicted *not* to change because of the treatment but is expected to respond to some or all of the contextually important internal validity threats in the same way as the target outcome" (p. 509). There is no limit on the number of nonequivalent DVs, but the main DV (the target outcome) and the nonequivalent DV should measure similar manifest or latent constructs (Coryn & Hobson, 2011). A nonequivalent DV can be added to many types of designs, such as multi-component. However, often the addition is to a one–group pretest–posttest design. Here, the main and nonequivalent DV are measured at the same time, pre and post intervention.

The results of the change on the main DV can be compared against the results of the change in the nonequivalent DV (i.e., pattern matching). Since both DVs are expected to be affected by the same contextual factors, participants act as their own control/comparison. For example, if an effect of maturation or history was present on the main DV, then it would also be expected on the nonequivalent DV. If the main DV changes over time, but there is small or no meaningful change in the nonequivalent DV, then this is evidence against threats to internal validity (such as maturation and history) as the cause of the change in the main DV. Thus, one has a set of alternative causal hypotheses, with the more complex the predicted pattern (e.g., if multiple nonequivalent variables are used or there are multiple differently timed interventions), the better the evidence that the cause of the change in the main DV was not due some common threats to interval validity.

An example of a nonequivalent DV design is Mitchell and Begeny (2014). In this study, a reading fluency intervention, the Helping Early Literacy with Practice Strategies (HELPS) program, was being evaluated. Several DVs were measured including reading efficiency, accuracy, and comprehension. Mitchell and Begeny (2014) hypothesized that if the HELPS program was effective, scores should increase over time. But in addition to the typical one–group pretest–posttest design, they also included several nonequivalent DVs: math achievement and phonetic decoding. While these are also academic abilities, and related to their main DVs, the HELPS program is *not* intended to enhance math achievement and phonetic decoding. Results indicated that there were medium to large effects in the expected direction on all the main DVs (Cohen's ds > 0.72 ps < .01), but they also found negative effects that were small to medium on the nonequivalent DVs of math achievement and phonetic decoding (Cohen's ds < −0.68 ps > .01). If Mitchell and Begeny (2014) had used the typical one–group pretest–posttest design, then increases in their reading outcomes could be attributed to effects such as maturation and not necessarily the HELPS program. However, since the targeted abilities increased greatly over time and the non-targeted abilities did not, there is better evidence that the HELPS program was effective. For other examples, see McKillip and Baldwin (1990); Schwab et al. (2019); and White (2000).

Applying the Nonequivalent Dependent Variable Design to Teaching Assessment

To apply the nonequivalent DV design, the instructor would measure the outcome/target of the course, which could be skill or knowledge. This would be the main DV. Next, for the nonequivalent DV, an instructor measures knowledge or skill that is not the target of the course and therefore predicted not to change. The instructor would measure both DVs at the beginning of the semester and at the end of the semester. For example, in a clinical psychology course, the main DV might be knowledge of clinical psychology assessed via 25 multiple choice items. The nonequivalent DV might be knowledge of social psychology assessed via 25 multiple choice items. Knowledge of social psychology is not intended to increase over the semester but should respond to the same contextually relevant factors as clinical psychology knowledge. Using this as an example, let us assume we calculate for each person, percentage scored, by averaging the 25 main items and multiplying by 100. This would also be done for the nonequivalent items. Therefore, there would be four variables, percentage scored on the main clinical psychology items at the beginning of the semester (Mpre), percentage scored on the main items at the end of the semester (Mpost), percentage scored on the nonequivalent social psychology items at the beginning of the semester (NEpre), percentage scored on the nonequivalent items at the end of the semester (NEpost). If there is an increase in Mpost compared to Mpre, but there is small or no meaningful change in NEpost compared to NEpre, then this is evidence of learning and not maturation, history, and testing effects. An example of this is provided in Figure 1, where data was simulated based on a large effect of the main DV and there was no change in the nonequivalent DV (see Reynolds, 2021 for R code).

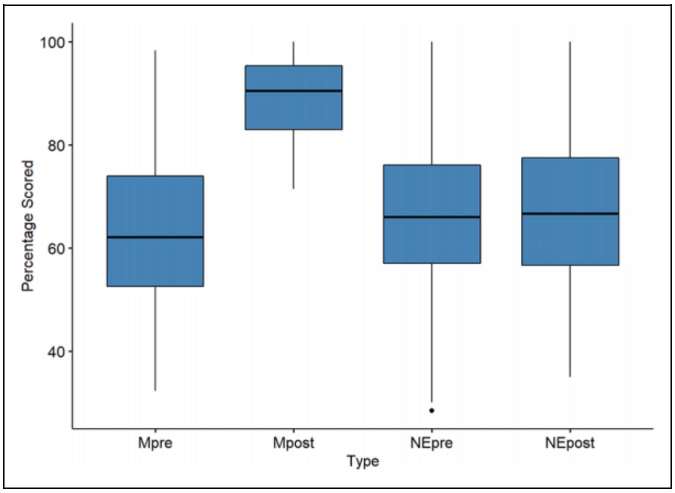

Figure 1. Boxplots from a nonequivalent dependent variable simulation. *Note*. Mpre = main dependent variable (pre), Mpost = Main dependent variable (post), NEpre = nonequivalent dependent variable (pre), and NEpost = nonequivalent dependent variable (post).

If the increase in Mpost compared to Mpre was due to a maturation effect such as change in study habits, a history effect such as stress, or a testing effect because they have been exposed to those items previously, then it would be likely that there would be an increase in NEpost compared to NEpre. Thus, an effect on the main DV and not on the nonequivalent DV is better evidence of a true effect of learning than the one–group pretest–posttest design. This does not rule out these effects; however, it provides evidence against them and strengthens the argument that changes on the main DV is less likely due to some threats to internal validity.

Choosing the Nonequivalent Dependent Variables

In choosing the type of nonequivalent DV, instructors should consider that it should be related to the main DV, but not overlap. For example, in an algebra course, if the main DV was knowledge in algebra, a nonequivalent DV might be knowledge in geometry. Algebra knowledge and geometry knowledge are both mathematical, but basic geometry does not necessarily require algebra to solve (e.g., finding the third angle in a two-dimensional triangle). The course in algebra is not intended to increase knowledge in geometry, yet it is likely that geometry knowledge is influenced by the same contextual factors.

In Mitchell and Begeny (2014), both the main DVs and the nonequivalent DVs were academic abilities. Therefore, both main DVs and the nonequivalent DVs arguably measured the same latent construct (in this case, general academic ability). But the HELPS intervention was not designed to increase math achievement or phonetic decoding, and while they are related with other academic abilities, math achievement does not cause reading abilities to increase.

The nonequivalent DV should also be dynamic (i.e., changeable). In Mitchell and Begeny (2014), it was possible for students to increase their math achievement or phonetic decoding abilities. If the nonequivalent DV is completely unrelated or not possible to change over time, then its inclusion is meaningless. As discussed earlier, the nonequivalent DV must be influenced by similar contextual factors as the main DV.

Nonequivalent dependent variables in psychology. An appropriate nonequivalent DV in a psychology course will depend on the specific course; however, knowledge of another subfield in psychology is likely reasonable. If the course being assessed was research methodology or statistics, then the nonequivalent DV could be knowledge in social, cognitive, physiological, evolutionary, developmental, or clinical psychology. Other psychology courses, such as forensic psychology, may pose a greater challenge.

Forensic psychology is a highly multidisciplinary sub-field in psychology with research from cognitive (e.g., eyewitness memory), developmental (e.g., child maltreatment), clinical (e.g., insanity defense), and social psychology (e.g., interrogation) often discussed. Knowledge in another sub-field of psychology might still be a reasonable nonequivalent DV; however, the instructor should carefully consider the degree of overlap with the assessed course.

The instructor could add multiple nonequivalent variables. For example, if the course being assessed was developmental psychology, one nonequivalent DV could be knowledge in statistics and another nonequivalent DV could be social psychology knowledge. The instructor might hypothesize a very small performance change in knowledge of statistics, a small to medium change in social psychology knowledge, and a large increase in developmental psychology knowledge. To the extent this is the case, the instructor now has even stronger evidence as to the cause of the change in developmental psychology knowledge.

Additional Considerations

To control for order effects, the nonequivalent and main items should have their answers appear in random order and the order of the type of item (main vs. nonequivalent) should also be randomized. Instructors would therefore benefit from using computers to administer the assessment rather than paper and pencil. The nonequivalent items should also be of similar form as the main items. For example, if the main items tend to be application items, then the nonequivalent items should also be mostly application items, to control for an effect of item level and/or type. Lastly, all items should be somewhat difficult, to avoid a ceiling effect, but difficult enough to avoid floor effects.

The focus here has been on assessing knowledge via multiple choice items, which is advantageous because they can be answered quickly, graded automatically, and graded more objectively. In using the nonequivalent DV approach, the items could be essay, fill in the blank, true/false, or other types. The instructor would only need to use the same types of items for the main DV and the nonequivalent DV. If an instructor used, for example, 20 multiple choice items and 5 short answer essay items for the main DV, then the nonequivalent DV should also have 20 multiple choice items and 5 short answer essay items. This is to reduce the probability of an item type confound.

Lastly, in addition to knowledge, skills can be assessed. For example, if writing were an emphasis in the course, such as in research methodology, a main DV could involve students writing one or more response papers which could be assessed for precision and persuasiveness. The instructor would carefully consider what writing skill is most emphasized. Conversely, for the nonequivalent DV, the instructor would carefully consider the writing skill least emphasized. For example, in a research methodology course the same response paper that is assessed for persuasiveness, which would constitute the main DV, could be assessed for grammar, sentence construction, and/or creativity, which would be the nonequivalent DV (this will depend on the course and instructor). Student's scores on this nonequivalent DV might increase slightly over the semester; however, if an instructor does not emphasize these skills, then the difference should be small, as compared to the change in scores on persuasiveness. In this case, a major advantage is that the response paper(s) that is being assessed is the same for the main DV as the nonequivalent DV. Alternatively, an instructor could have multiple response papers, for example, some assessed for persuasiveness and others assessed for creativity. The instructor should ensure that each prompt is of similar difficulty.

To be clear, the same data analytic strategy can be followed for skill assessment. In this case, Mpre might be a 0–100 score on persuasiveness of the writing for the pre-test. Mpost would then be the score on persuasiveness to the same prompt on the post-test. NEpre might be a 0-100 score on grammar in the response on the pre-test. NEpost would then be the score on grammar on the post-test. If students wrote multiple response papers the instructor could average them. If a large change is detected in Mpre compared to Mpost and a small or no change is detected in NEpre compared to NEpost, then the instructor has better evidence that students learned the specific skills taught in the course. On the other hand, for example, the change was small for both the main DV and the nonequivalent DV, the instructor has evidence that the skills that were a priority (e.g., persuasiveness) were not emphasized enough or taught effectively.

Instead of writing, the skill might be quantitative reasoning. For example, in a statistics course, if interpreting *p* values were emphasized, the main DV might involve giving students tables of results (e.g., an ANOVA table) and students would have to interpret the relevant information in a written response. If frequentists statistics were the focus of the course, a nonequivalent DV might involve including a Bayes Factor in the table which students would have to interpret. On the other hand, if the course emphasized Bayesian statistics, the main DV might involve correct interpretation of Bayes Factors in the table and the nonequivalent DV might be interpreting *p* values.

Design Comparison

This nonequivalent DV approach, compared to the one–group pretest–posttest design with alternative forms, allows the instructor to make a stronger case for meaningful change in the main DV. This is because the nonequivalent DV design has higher internal validity. The nonequivalent DV design is also less logistically challenging. In the nonequivalent DV design, all students answer both sets of items at the same time. Compare this to the alternative forms design which requires giving different students different versions at different times. Therefore, the nonequivalent DV design not only allows for stronger conclusions about the observed effects but is easier to implement.

Data Analysis

One potential issue with implementing new approaches is that some might not feel the extra work is justified. Indeed, research indicates that making a strategy or intervention accessible increases its use (e.g., Zhang et al., 2016). To address this, an R script has been created for analyzing these data (see Reynolds, 2021; Assessment script 1). Using their program of choice, instructors need to match up student's data from pre to post as would typically be done. Next, instructors should make their dependent variables, for example by averaging or calculating a percentage, for the appropriate items. There should be at least five variables: The student's average on the pre-test for the main items (label this "Mpre"), average on the post-test for the main items (label this "Mpost"), average on the pre-test for the nonequivalent items (label this "NEpre"), average on the post-test for the nonequivalent items (label this "NEpost"), and an identification number for each student (e.g., 1-40 if there are 40 paired scores; label this "id"). The name of this datafile should be "d". Instructors should then import the data into R (R studio is recommended), run the R script provided, and all analyses will be generated. Instructors can also highlight just the sections they wish to run. Included are descriptive statistics for all DVs, two paired boxplots with labels, and a plot with all four variables plotted together as boxplots. For those interested in frequentist hypothesis testing, there is data screening, two dependent samples *t* tests (one test for the main DV and one test for the nonequivalent DV), and Hedge's g (with CIs). For those interested in Bayesian analyses, included is a Bayesian version of the *t* test, Bayesian estimation supersedes the *t* test (BEST), which outperforms the frequentist test (see Bååth, 2014; Kruschke, 2013), plots from the BEST analysis, and two Bayes Factors.

There are many statistical approaches that would be useful for analyzing such data. The traditional *t* tests have been included, as most psychologists have basic knowledge of *t* tests. However, the included Bayesian analyses are preferred. While psychology instructors may be more familiar with SPSS, R software is free and, using the script provided, all analyses are generated easily and require very little preexisting knowledge of R. To facilitate these analyses for instructors who may be unfamiliar with R, included in the link of supplementary materials is a video tutorial explaining how to use the script to analyze example data. A second R script has been provided if two nonequivalent DVs are used (Assessment script 2). The assessment scripts can also be used if skills are assessed, as in the writing example given earlier.

Limitations

There are clear advantages of using a nonequivalent DV design over the traditional one–group pretest–posttest design. However, the nonequivalent DV design has several limitations. The major limitation is that since its purpose, as described here, is to evaluate learning over the semester, it would not be useful for evaluating incremental changes in the students. To assess incremental changes, the embedded outcomes approach is useful (see McCarthy et al., 2011); however, it has many methodological issues. If students perform well on an exam or a paper, it does not necessarily demonstrate learning or effective teaching, the test items may be too easy or essays graded too leniently, it could be due to maturation, motivation, or a history effect. Thus, the embedded outcomes approach has none of the methodological strengths of nonequivalent DV design and would be a complementary, rather than an alternative, method for learning assessment.

While there is an increase in interval validity when using a nonequivalent DV design, the design does not allow an instructor to rule out all alternative hypotheses. An increase on the main DV could be genuinely attributed to learning; however, the learning may not be a product of teaching effectiveness but some other cause. There are many causes of learning, with teaching effectiveness being just one. It is also possible that some other cause of learning might interact with a maturation, a history, or a testing effect. Depending on the nature of the interaction, it could produce an increase in the main DV but not in the nonequivalent DV. If there is missing data (i.e., some students do not do either the pre or post-test) and this causes a selection bias, the effect of learning is biased. It is then possible for a selection by maturation interaction to bias the effect. Therefore, the nonequivalent DV design does not rule out these types of effects.

Conclusions

Given the increasing awareness of the issues with SET, instructors may wish to use learning outcomes approaches to improve their teaching and demonstrate effective teaching for tenure purposes. The nonequivalent DV design has considerable advantages over the one–group pretest–posttest design and the alternative forms design, while requiring only minimal effort. There are still issues with this design, but it can allow the instructor to make a stronger case that learning was the cause of the change on the main DV. Furthermore, the design is flexible, allowing the instructor to incorporate multiple main DVs and/or multiple nonequivalent DVs, which could be knowledge or skills.

Declaration of Conflicting Interests

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

Funding

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

References

Aczel, B., Palfi, B., & Szaszi, B. (2017). Estimating the evidential value of significant results in psychological science. *PLOS ONE*, *12*(8), 1–8. https://doi.org/10.1371/journal.pone.0182651563279

Arthur, W., Jr., Tubré, T., Paul, D. S., & Edens, P. S. (2003). Teaching effectiveness: The relationship between reaction and learning evaluation criteria. *Educational Psychology*, *23*(3), 275–285. https://doi.org/10.1080/01443410320000060110

Bååth, R. (2014). Bayesian first aid: A package that implements Bayesian alternatives to the classical test functions in R. In *the Proceedings of User! The International R User Conference*.

Bartsch, R. A., Bittner, W. M. E., & Moreno, J. E., Jr. (2008). A design to improve internal validity of assessments of teaching demonstrations. *Teaching of Psychology*, *35*(4), 357–359. https://doi.org/10.1080/00986280802373809

Berk, R. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, *17*(1), 48–62.

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *Science Open Research*, 1–11. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Brough, P. (Ed.). (2019). *Advanced research methods for applied psychology: Design, analysis and reporting*. Routledge.

Coryn, C. L. S., & Hobson, K. A. (2011). Using nonequivalent dependent variables to reduce internal validity threats in quasi-experiments: Rationale, history, and examples from practice. *New Directions for Evaluation*, *131*, 31–39. https://doi.org/10.1002/ev.375

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education*, *45*(8), 1106–1120. https://doi.org/10.1080/02602938.2020.1724875

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org10.1053/j.seminhematol.2008.04.003

Hoorens, V., Dekkers, G., & Deschrijver, E. (2020). Gender bias in student evaluations of teaching: Students' self-affirmation reduces the bias by lowering evaluations of male professors. *Sex Roles: A Journal of Research*, *84*(1), 34–48. https://doi.org/10.1007/s11199-020-01148-8

Ioannidis, J. P. A. (2019). What have we (not) learnt from millions of scientific papers with p values? *American Statistician*, *73*(1), 20–25. https://doi.org/10.1080/00031305.2018.1447512

Kruschke, J. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. http://doi.org/10.1037/a0029146

Kruschke, J., & Liddell, T. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. http://doi.org/10.3758/s13423-016-1221-4

McCarthy, M. A., Niederjohn, D. M., & Bosack, T. N. (2011). Embedded assessment: A measure of student learning and teaching effectiveness. *Teaching of Psychology*, *38*(2), 78–82. https://doi.org/10.1177/0098628311401590

McKillip, J., & Baldwin, K. (1990). Evaluation of an STD education media campaign: A control construct design. *Evaluation Review*, *14*(4), 331–346. https://doi.org/10.1177/0193841X9001400401

McSweeny, J. A. (1978). Effects of response cost on the behavior of a million persons: Charging for directory assistance in Cincinnati. *Journal of Applied Behavior Analysis*, *11*, 47–51. https://doi.org/10.1901/jaba.1978.11-47

Mitchell, C., & Begeny, J. C. (2014). Improving student reading through parents' implementation of a structured reading program. *School Psychology Review*, *43*(1), 41–58. https://doi.org/10.1080/02796015.2014.12087453

Olivares, O. J. (2003). A conceptual and analytic critique of student ratings of teachers in the USA with implications for teacher effectiveness and student learning. *Teaching in Higher Education*, *8*(2), 233–245. https://doi.org/10.1080/1356251032000052465

Reynolds, J. J. (2021, May 9). *Improving the assessment of teaching effectiveness with the nonequivalent dependent variables approach*. https://doi.org/10.17605/OSF.IO/23MJP

Robertson, T. S., & Rossiter, J. R. (1976). Short-run advertising effects on children: A field study. *Journal of Marketing Research*, *13*, 68–70. https://doi.org/10.1177/002224377601300109

Schwab, J. R., Houchins, D. E., Peng, P., McKeown, D., Varjas, K., & Emerson, J. (2019). The effects of a multi-component intervention to increase math performance for students with EBD in alternative educational settings. *International Journal of Special Education*, *34*(1), 226–244.

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

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin*, *26*(11), 1329–1342. https://doi.org/10.1177/0146167200263002

Stark-Wroblewski, K., Ahlering, R., & Brill, F. (2007). Toward a more comprehensive approach to evaluating teaching effectiveness: Supplementing student evaluations of teaching with pre-post learning measures. *Assessment & Evaluation in Higher Education*, *32*(4), 403–415. https://doi.org/10.1080/02602930600898536

Szucs, D., & Ioannidis, D. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience*, *11*(390), 1–21. https://doi.org/10.3389/fnhum.2017.00390

Trochim, W. M., & Donnelly, J. P. (2016). *Research methods: The essential knowledge base* (2nd ed.). Cengage Learning.

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wagenmakers, E., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57. https://doi.org/10.3758/s13423-017-1343-3

White, M. D. (2000). Assessing the impact of administrative policy on use of deadly force by on- and off-duty police. *Evaluation Review*, *24*(3), 295–318. http://doi.org/10.1177/0193841X0002400303

Zhang, S., Zhang, M., Yu, X., & Ren, H. (2016). What keeps Chinese from recycling: Accessibility of recycling facilities and the behavior. *Resources, Conservation & Recycling*, *109*, 176–186. https://doi.org/10.1016/j.resconrec.2016.02.008