Design and Analysis in Educational Research:

ANOVA Design Case Studies for Teaching Race, Racism, and Black Lives Matter

Kamden K. Strunk, Mwarumba Mwavita, and Payton D. Hoover

Note: These case studies are intended as an online instructional supplement for Strunk, K., & Mwavita, M. (2020). *Design and Analysis in Educational Research: ANOVA Designs in IBM SPSS*. Routledge. The presentation of these cases follows the process taught in that book closely and supplements case studies published in the book. We recommend using these cases in conjunction with that textbook, though using them does not require the purchase or use of that textbook. Data files are available on the Routledge website at www.routledge.com/9781138361164

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

CONTENTS

INTRODUCTION TO THIS RESOURCE

This set of case studies and the accompanying data files are intended as a supplement to our statistical analysis textbook, which is available as either an SPSS or jamovi version through Routledge. As a result, these case studies do not include direct demonstration or instruction around the conceptual foundations of each design, how to work the SPSS or jamovi software, or specific steps used to produce appropriate output. All of those steps are covered in our statistics textbooks or can be learned from other sources. Our statistics textbooks contain case studies for every design they cover, and the content of those cases is wide-ranging, including equity research, educational psychology, educational leadership, teaching and learning, and other areas of educational research. During the Summer of 2020, as broad coalitions of social movements sought justice for the murder of Black people across the U.S. and many looked for ways to engage anti-racism in their teaching and research, we saw quantitative methods instructors and students searching for ways to engage. We hope that this set of case studies is useful in that effort.

We want to acknowledge at the beginning of this resource package that quantitative methods have a history of entanglement with racism, white supremacy, and eugenics. A number of excellent resources exist for understanding and coming to terms with that history and what it means for contemporary educational research. One result of that history, though, is that the published research on race and racism is often lacking, problematic, or explicitly racist. There is, of course, some high-quality quantitative scholarship on race and racism, but there is also a lot of work that is not. In selecting cases for this resource, one of the limitations has been that we needed to find examples of very specific ANOVA methods being used. Many of the published articles we located used more advanced methods (like HLM, SEM, regression) that are beyond the scope of our textbook. We tried to locate published articles that had relatively straightforward designs that could be useful for students learning a particular design and analysis, that were not especially problematic in their treatment of race and/or racism, and that provided enough detail for us to create simulated data sets.

On that note, while we discuss this in the full textbook, we here wish to make a short note about the data sets provided on the Routledge website. Those data sets are not 'real' data but are data sets that we have simulated to reproduce the results of the published articles. There are a few important facts about the simulation process that affect how those data work with case studies. First, the process by which we simulated the data produces a normal distribution. Thus, while we discuss the assumption of normality for each case study, it cannot be meaningfully tested in the simulated data, which will always be nearly perfectly normally

distributed. Second, the data simulation process does not perfectly reproduce the results of within-subjects or repeated measures designs, so while those data correctly reproduce the basic direction and nature of the relationships or differences, they will not perfectly match the published results. Our writing in this resource set, though, will reflect the values calculated from the simulated data, they just will not perfectly match what is in the published articles, which we hope those using this resource will read.

This resource is being made available free of charge by Routledge, the publisher of our statistics textbooks. While we, of course, hope that people will find those books useful and will use this resource as an additional set of cases for the books, we know there are many other statistics textbooks available. The cases throughout this resource are analyzed and written up in a format we teach in our textbooks, but that is likely fairly close to most formats in other books. As a result, we believe this resource can be useful in teaching about race and racism in educational research for people using a wide range of instructional texts. For all of the included case studies, datasets are available on the Routledge website for our textbooks as a free download, as well.

INDEPENDENT SAMPLES *t*-TEST CASE STUDIES

In this chapter, we will present several examples of published research that used an independent samples *t*-test. For each sample, we encourage you to:

1. Use your library resources to find the original, published article. Read that article and look for how they use and talk about the *t*-test.
2. Visit this book's online resources and download the datasets that accompany this chapter. Each dataset is simulated to reproduce the outcomes of the published research. (Note: the online datasets are not real human subjects data but have been simulated to match the characteristics of the published work.)
3. Follow along with each step of the analysis, comparing your own results with what we provide in this chapter. This will help cement your understanding of how to use the analysis.

CASE STUDY 1: IMPLICIT BIAS IN SCHOOL DISCIPLINE

Marcucci, O. (2020). Implicit bias in the era of social desirability: Understanding antiblackness in rehabilitative and punitive school discipline. *Urban Review: Issues and Ideas in Public Education*, *52*(1), 47–74. https://doi.org/10.1007/s11256-019-00512-7

In this paper, the author examined if implicit bias impacted seven disciplinary decisions, from punitive to rehabilitative disciplines. The researcher wanted to see if perceptions between teachers would vary depending on the race of the student to be disciplined. A survey was administered to a sample of 287 teachers, randomly assigning them to a racial priming condition. Participants were either in the White racial prime condition (i.e., told the student was White) or the Black racial prime condition (i.e., told the student was Black).

Research Questions.

The researcher was interested in examining the following two questions:
Research Question 1: Are there differences in teachers' likelihood for all disciplinary decisions based on whether or not they were placed in the White or Black racial priming condition?

Research Question 2: Are there differences in teachers' philosophical support for exclusionary discipline based on whether or not they were placed in the White or Black racial priming condition?

Hypotheses.

The author hypothesized the following related to teachers' likelihood for all disciplinary decisions:

H_0: There was no significant difference between those in the White racial priming condition and those in the Black racial priming condition on perceptions of likelihood for all disciplinary decisions.

H_1: There was a significant difference between those in the White racial priming condition and those in the Black racial priming condition on perceptions of likelihood for all disciplinary decisions.

The author hypothesized the following related to teachers' philosophical support for exclusionary discipline:

H_0: There was no significant difference between those in the White racial priming condition and those in the Black racial priming condition on perceptions of philosophical support for exclusionary discipline.

H_1: There was a significant difference between those in the White racial priming condition and those in the Black racial priming condition on perceptions of philosophical support for exclusionary discipline.

Variables Being Measured.

There were two hypotheses that the research tested: teachers' likelihood for all disciplinary decisions and teachers' philosophical support for exclusionary discipline. For the hypothesis regarding teachers' likelihood for all disciplinary decisions, the dependent variables (DV) were the seven distinct types of disciplinary decisions (7 item survey). The author noted that the survey responses were on a 7-point Likert scale and were re-coded during their analyses to indicate stronger agreement for reliability purposes. For the hypothesis regarding teachers' philosophical support for exclusionary discipline, the dependent variable was a disciplinary philosophy subscale (3 items). The author noted that the three items on this subscale has a Cronbach's alpha of .78, ensuring reliability.

Conducting the Analysis.

1. What test did they use, and why?

The author used an independent samples *t*-test to determine if racial priming condition correlated with teachers' likelihood for all disciplinary decisions and their philosophical support for exclusionary discipline.

2. What are the assumptions of the test? Were they met in this case?
    a. Level of Measurement for the Dependent Variable is Interval or Ratio
       The two dependent variables were teachers' likelihood for all disciplinary decisions and philosophical support for exclusionary discipline. Both DVs are on interval scales.
    b. Normality of the Dependent Variable
       The author assumed normality of the dependent, therefore they did not share the skew and kurtosis values.
    c. Observations are Independent
       The observations were independent from one another through the use of random assignment and the survey being completed individually online.
    d. Random Sampling and Assignment
       The author used a snowball sample of teachers online.
       Participants were randomly assigned to a racial priming condition.
    e. Homogeneity of variance
       The assumption of homogeneity of variance was met for both teachers' likelihood for all disciplinary decisions ($F = .013$, $p = .908$) and philosophical support for exclusionary discipline ($F = .797$, $p = .373$).
3. What was the result of that test?

There was a significant difference in teachers' likelihood for all disciplinary decisions ($t_{284.868} = 4.184$, $p < .001$) and in teachers' philosophical support for exclusionary discipline ($t_{282.501} = 2.517$, $p = .012$).

4. What was the effect size, and how is it interpreted?

The authors reported Cohen's *d*. However, we could calculate omega squared for each test:

$$Likelihood: \omega^2 = \frac{t^2 - 1}{t^2 + N_X + N_Y - 1} = \frac{4.184^2 - 1}{4.184^2 + 144 + 143 - 1}$$
$$= \frac{17.506 - 1}{17.506 + 144 + 143 - 1} = \frac{16.506}{303.506} = .054$$
$$Support: \omega^2 = \frac{t^2 - 1}{t^2 + N_X + N_Y - 1} = \frac{2.517^2 - 1}{2.517^2 + 144 + 143 - 1}$$
$$= \frac{6.335 - 1}{6.335 + 144 + 143 - 1} = \frac{5.335}{292.335} = .018$$

From these calculations, we can determine that about 5% of the variance in teachers' likelihood for all disciplinary decisions ($\omega^2 = .054$) and about 2% of the variance in teachers' philosophical support for exclusionary discipline ($\omega^2 = .018$) was explained by the racial priming condition they were in.

5. What is the pattern of group differences?

Those placed in the White racial prime condition (M = 5.250, SD = 2.050) had higher ratings of likelihood for disciplinary decisions when compared to those placed in the Black racial prime condition (M = 4.230, SD = 2.080). Similarly, when examining teachers' philosophical support for exclusionary discipline, those in the White prime condition (M = 3.760, SD = 1.660) had higher ratings on the items than those in the Black prime condition (M = 3.290, SD = 1.500).

Write up.

Results

We used an independent samples *t*-test to determine if there were mean differences in a) teachers' likelihood for all disciplinary decisions and b) their philosophical support for exclusionary discipline based on racial priming condition. There was a significant difference in teachers' likelihood for all disciplinary decisions ($t_{284.868}$ = 4.184, p < .001) and in teachers' philosophical support for exclusionary discipline ($t_{282.501}$ = 2.517, p = .012). About 5% of the variance in teachers' likelihood for all disciplinary decisions was explained by the racial priming condition they were in. About 2% of the variance in teachers' philosophical support for exclusionary discipline was explained by the racial priming condition they were in. Those placed in the White racial prime condition (M = 5.250, SD = 2.050) had higher ratings of likelihood for disciplinary decisions when compared to those placed in the Black racial prime condition (M = 4.230, SD = 2.080). Similarly, when examining teachers' philosophical support for exclusionary discipline, those in the White prime condition (M = 3.760, SD = 1.660) had higher ratings on the items than those in the Black prime condition (M = 3.290, SD = 1.500).

CASE STUDY 2: LITERACY ACHIEVEMENT IN ELEMENTARY SCHOOL

Matthews, J. S., Kizzie, K. T., Rowley, S. J., & Cortina, K. (2010). African Americans and boys: Understanding the literacy gap, tracing academic trajectories, and evaluating the role of learning-related skills. *Journal of Educational Psychology*, *102*(3), 757. https://doi.org/10.1037/a0019616

In this study, the authors utilized the Early Childhood Longitudinal Study-Kindergarten (ECLS-K) to examine race and gender-based differences in academic achievement among African American and White children in kindergarten to 5th grade. The author's focal point was the literacy achievement of African American boys, while also considering learning-related skills (LRS) and other external social factors. Furthermore, the ECLS-K provided the authors with data from the 1998-99 cohort of U.S. public and private schools which consisted of a total of 12,385 kindergarten children.

Research Questions.

The researchers were interested in determining:
Are there differences in literacy achievement scores between African American boys, White boys, and African American girls from kindergarten to fifth grade?

Hypotheses.

The authors hypothesized the following:

H_0: There was no significant difference in literacy gaps between both African American boys and White boys and African American girls from kindergarten through fifth grade.

H_1: There was a significant difference in literacy gaps between both African American boys and White boys and African American girls from kindergarten through fifth grade.

Variables Being Measured.

The authors utilized the item response theory (IRT) to assess the literacy achievement for the sample of students. This study measured three areas of children's classroom behavior (those variables included: Learning approaches (six items); Externalizing problem behaviors (five items); and Interpersonal skills (five items). The other variables included: the influence of SES, home literacy environment and literacy achievement from kindergarten to 5th grade (which were measured at 6 different point in times). To measure these items, the authors utilized the Social Skills Rating Scale and noted that the children were rated on a 4-point Likert scale. The authors confirmed reliability by noting that the literacy achievement ratings (basic

and advanced) have a Cronbach's alpha ranging from .75 to .88 from kindergarten to the fifth grade. The reliability of the overall IRT literacy ability estimates (theta) were also strong, ranging from .91 to .96 from kindergarten to the fifth grade.

Conducting the Analysis.

1. What test did they use, and why?

The authors used independent-samples *t*-tests to determine if there was an increase in literacy gaps between African American boys, White boys, and African American girls from kindergarten through fifth grade.

2. What are the assumptions of the test? Were they met in this case?
   a. Level of Measurement for the Dependent Variable is Interval or Ratio
      The dependent variables were literacy achievement scores for kindergarten and fifth grade, both of which were on an interval scale.
   b. Normality of the Dependent Variable
      In most cases, when this assumption is met, the article will not report normality statistics. Because the authors did not report about normality, we must infer this assumption was met. However, normally researchers would evaluate this assumption before running the analysis in which we would check for normality using skewness and kurtosis statistics. We found that literacy achievement in fifth grade was negatively skewed (*skewness* = -.143, *SE* = .031) and leptokurtic (*kurtosis* = .161, *SE* = .061).
   c. Observations are Independent
      The authors did not report any evidence of dependence.
   d. Random Sampling and Assignment
      The authors did not use random sampling or assignment because they used data previously collected (secondary data) from ECLS-K and NCES.
   e. Homogeneity of variance
      The authors did not report whether or not the assumption for homogeneity of variance was met or not. When running the analysis with the simulated data, we found that the assumption was not met for literacy achievement scores in kindergarten ($F = 100.68$, $p < .001$), nor for the scores in fifth grade ($F = 52.763$, $p < .001$).
3. What was the result of that test?

There was a significant difference in literacy achievement in kindergarten between African American students and White students ($t_{2420.053} = -15.261$, $p < .001$) and literacy achievement in fifth grade between African American students and White students ($t_{1728.924} = -25.423$, $p < .001$).

4. What was the effect size, and how is it interpreted?

The authors reported Cohen's *d*. However, we could calculate omega squared for each test:

$$Literacy K: \omega^2 = \frac{t^2 - 1}{t^2 + N_X + N_Y - 1}$$
$$= \frac{-15.261^2 - 1}{-15.261^2 + 1257 + 5086 - 1}$$
$$= \frac{232.898 - 1}{232.898 + 1257 + 5086 - 1} = \frac{231.898}{6574.898} = .035$$
$$Literacy 5: \omega^2 = \frac{t^2 - 1}{t^2 + N_X + N_Y - 1}$$
$$= \frac{-25.423^2 - 1}{-25.423^2 + 1257 + 5086 - 1}$$
$$= \frac{646.329 - 1}{646.329 + 1257 + 5086 - 1} = \frac{645.329}{6988.329} = .092$$

From these calculations, we can determine that about 4% of the variance in literacy achievement scores in kindergarten (ω^2 = .035) and about 9% of the variance in literacy achievement scores in fifth grade (ω^2 = .092) was explained by race.

5. What is the pattern of group differences?

In kindergarten, African American students (*M* = 26.100, *SD* = 8.000) scored lower in literacy achievement than White students (*M* = 30.200, *SD* = 10.400). In fifth grade, the pattern was the same and differences between African American students (*M* = 124.000, *SD* = 25.700) and White students (*M* = 144.000, *SD* = 21.800) increased from kindergarten.

Write up.

Results

The authors used an independent-samples *t*-tests to determine if there was an difference in literacy gaps between African American boys, White boys, and African American girls from kindergarten through fifth grade. Some of the assumptions for this test were not normal. Literacy achievement in fifth grade was negatively skewed (*skewness* = -.143, *SE* = .031) and leptokurtic (*kurtosis* = .161, *SE* = .061)**.** The assumption for homogeneity of variance was not met for literacy achievement in kindergarten (*F* = 100.68, *p* < .001) nor for literacy achievement

in fifth grade (F = 52.763, p < .001). There was a significant difference in in literacy achievement in kindergarten between African American students and White students ($t_{2420.053}$ = -15.261, p < .001) and literacy achievement in fifth grade between African American students and White students ($t_{1728.924}$ = -25.423, p < .001). About 4% of the variance in literacy achievement scores in kindergarten was explained by race. About 9% of the variance in literacy achievement scores in fifth grade was explained by race. In kindergarten, African American students (M = 26.100, SD = 8.000) scored lower in literacy achievement than White students (M = 30.200, SD = 10.400). In fifth grade, the pattern was the same and differences between African American students (M = 124.000, SD = 25.700) and White students (M = 144.000, SD = 21.800) increased from kindergarten.

CASE STUDY 3: EFFECTIVENESS OF DIVERSITY TRAINING

Saleh, M. F., Anngela-Cole, L., & Boateng, A. (2011). Effectiveness of diversity infusion modules on students' attitudes, behavior, and knowledge. *Journal of Ethnic & Cultural Diversity in Social Work*, *20*(3), 240–257. https://doi.org/10.1080/15313204.2011.594995

In this paper, the authors examined the effectiveness of diversity infusion modules for students from a predominantly white university to participate in. These infusion models focused on racism, power, white privilege, oppression, and health disparities. The participants involved in the study were social work and allied helping professional students from a land grant university. The social work students participated in the infusion models, while the other students did not (i.e., referred to as the comparison group). The researchers administered measures testing participants attitudes, knowledge, or behaviors about diversity.

Research Questions.

The researchers were interested in determining:
Did participation in the diversity infusion modules make a difference in attitudes, knowledge, or behaviors about diversity for social work students?

Hypotheses.

The authors hypothesized the following related to knowledge:

H_0: There was no significant difference between social work students and the comparison group on knowledge test scores.

H_1: There was a significant difference between social work students and the comparison group on knowledge test scores.

The authors hypothesized the following related to the MRI (Multiculturally Responsible Index) test:

H_0: There was no significant difference between social work students and the comparison group on MRI test scores.

H_1: There was a difference between social work students and the comparison group on MRI test scores.

Variables Being Measured.

There were two hypotheses that the research tested: whether or not there was a difference between knowledge and MRI scores for social work students and the comparison group. Knowledge was measured using a ten-item survey assessing the participants knowledge on topics surrounding diversity, race, power, oppression, privilege, etc. The MRI measure

consisted of fourteen items assessing attitudes and behavior, with eight items focused on attitudes and the other six items focused on behaviors.

Conducting the Analysis.

1. What test did they use, and why?

The authors used an independent samples *t*-test to determine if there were differences in the MRI and knowledge tests between social work students who received the intervention and the non-social work students who did not receive the intervention.

2. What are the assumptions of the test? Were they met in this case?
   a. Level of Measurement for the Dependent Variable is Interval or Ratio
      The dependent variables were knowledge and MRI scores, both of which were measured using interval scales.
   b. Normality of the Dependent Variable
      Because the authors did not report about normality, we must infer this assumption was met.[1] However, normally researchers would evaluate this assumption before running the analysis in which we would check for normality using skewness and kurtosis statistics. Using the simulated data, we found that MRI scores were negatively skewed (*skewness* = -.412, *SE* = .195) and normal for kurtosis (*kurtosis* = .304, *SE* = .389), whereas knowledge scores were normal for skew (*skewness* = -.265, *SE* = .195) and leptokurtic (*kurtosis* = 1.063, *SE* = .389).
   c. Observations are Independent
      The authors did not report any known evidence of dependence.
   d. Random Sampling and Assignment
      The authors used a convenience sample comprised of students from a large, western land grant institution in the U.S.
      There was no possibility of random assignment because it was based on the students' areas of study.
   e. Homogeneity of variance
      The authors did not report whether or not the assumption for homogeneity of variance was met or not. When running the analysis with the simulated

[1] The process by which we simulate data for these case studies results in data that are almost perfectly normally distributed. Remember that the example datasets on the online resources are not actual human subjects data, but simulated data to reproduce the outcomes of the case study articles. If you decide to run the tests for normality for practice on these datasets, keep in mind they will be nearly perfect due to the manner in which we have simulated those data.

data, we found that the assumption was not met for MRI scores ($F = 7.443$, $p = .007$). However, the assumption of homogeneity of variance was met for knowledge scores ($F = 1.776$, $p = .185$).

3. What was the result of that test?

There was a significant difference in MRI scores between social work students and the comparison group ($t_{75.119} = 3.111$, $p = .003$), as well as a significant difference in knowledge scores between groups ($t_{152} = 7.031$, $p < .001$).

4. What was the effect size, and how is it interpreted?

The authors reported eta squared (η^2). However, we could calculate omega squared for each test:

$$Knowledge: \omega^2 = \frac{t^2 - 1}{t^2 + N_X + N_Y - 1} = \frac{7.031^2 - 1}{7.031^2 + 104 + 50 - 1}$$
$$= \frac{49.435 - 1}{49.435 + 104 + 50 - 1} = \frac{48.435}{202.435} = .239$$
$$MRI: \omega^2 = \frac{t^2 - 1}{t^2 + N_X + N_Y - 1} = \frac{3.111^2 - 1}{3.111^2 + 104 + 50 - 1}$$
$$= \frac{9.678 - 1}{9.678 + 104 + 50 - 1} = \frac{8.678}{162.678} = .053$$

From these calculations, we can determine that about 24% of the variance in knowledge ($\omega^2 = .239$) and 5% of the variance in MRI scores ($\omega^2 = .053$) was explained by whether or not the intervention was received.

5. What is the pattern of group differences?

For the MRI measure, the social work students ($M = 58.760$, $SD = 8.900$) scored higher than the comparison group ($M = 52.760$, $SD = 12.160$). The same pattern was observed for knowledge scores, with the social work students ($M = 8.510$, $SD = 1.960$) scoring higher than the comparison group ($M = 6.020$, $SD = 2.250$).

Write up.

Results

The authors used an independent samples *t*-test to determine if there were differences in the MRI and knowledge tests between social work students who received the intervention and the non-social work students who did not receive the intervention. Some of the assumptions of the test were not met. MRI scores were negatively skewed (*skewness* = -.412, *SE* = .195) and normal for kurtosis (*kurtosis*

= .304, *SE* = .389), whereas the knowledge scores were normal for skew (*skewness* = -.265, *SE* = .195) and leptokurtic (*kurtosis* = 1.063, *SE* = .389). In addition to this, the assumption of homogeneity of variance was not met for the MRI variable ($F = 7.443$, $p = .007$). There was a significant difference in MRI scores between social work students and the comparison group ($t_{75.119} = 3.111$, $p = .003$), as well as a significant difference in knowledge scores between groups ($t_{152} = 7.031$, $p < .001$). About 5% of the variance in MRI scores was explained by whether or not the intervention was received. About 24% of the variance in knowledge was explained by whether or not the intervention was received. For the MRI measure, the social work students ($M = 58.760$, $SD = 8.900$) scored higher than the comparison group ($M = 52.760$, $SD = 12.160$). The same pattern was observed for knowledge scores, with the social work students ($M = 8.510$, $SD = 1.960$) scoring higher than the comparison group ($M = 6.020$, $SD = 2.250$).

Now, compare this version, which follows the format we suggested in Chapter 6 of the textbook, to the published version. What is different? Why is it different? Notice that, in the full article, the *t*-tests are just one step among several analyses the authors used. Using the *t*-test in conjunction with other analyses, as these authors have done, results in some changes in how the test is explained and presented.

ONE-WAY ANOVA CASE STUDIES

In Chapter 8 of the textbook, we explored the one-way ANOVA using a made-up example and some fabricated data. In this section, we will present several examples of published research that used the one-way ANOVA. For each sample, we encourage you to:

1) Use your library resources to find the original, published article. Read that article and look for how they use and talk about the *t*-test.
2) Visit this book's online resources and download the datasets that accompany this chapter. Each dataset is simulated to reproduce the outcomes of the published research. (Note: the online datasets are not real human subjects data but have been simulated to match the characteristics of the published work.)
3) Follow along with each step of the analysis, comparing your own results with what we provide in this chapter. This will help cement your understanding of how to use the analysis.

CASE STUDY 1: OUTNESS AMONG QUEER AND TRANS STUDENTS OF COLOR

Garvey, J. C., Mobley, S. D., Summerville, K. S., & Moore, G. T. (2019). Queer and trans* students of Color: Navigating identity disclosure and college contexts. *Journal of Higher Education, 90*(1), 150-178. https://doi.org/10.1080/00221546.2018.1449081

In this paper, the authors explore outness and identity disclosure for queer and trans students of Color in college contexts. Their data came from an online survey that was distributed via email and social media, targeting college graduates who were LGBTQ. For this paper, their analysis focused on students of Color. The full study involves quantitative analyses as well as qualitative analysis of narrative data from participants. The focus of the ANOVA analysis was to determine if there were difference in levels of outness between various racial identities among queer and trans students of Color. In the qualitative analysis, the authors seek to understand those differences and the social and institutional conditions that surround decisions about gender and sexual identity disclosure for students of Color.

Research Questions.

We focus here on one research question that the authors answered using the one-way ANOVA:

1. Were there significant differences in levels of outness between students of Color based on racial identity (Asian/Pacific Islander, Black/African American, Latinx, and multiracial)?

Hypotheses.

The authors hypothesized the following:

H_0: There was no statistically significant difference in outness among Asian/Pacific Islander, Black/African American, Latinx, and multiracial students. ($M_{Asian} = M_{Black} = M_{Latinx} = M_{Multiracial}$)

H_1: There was a statistically significant difference in outness among Asian/Pacific Islander, Black/African American, Latinx, and multiracial students. ($M_{Asian} \neq M_{Black} \neq M_{Latinx} \neq M_{Multiracial}$).

Variables Being Measured.

The dependent variable in this study was outness, which is an effort to measure the extent to which students disclosed their sexual and gender identities in college. The authors report using a nine-item Likert-type scale adapted from an Outness Inventory. The outness scores were calculated as an average of the nine items, and then converted to standard scores (z-scores) to aid in interpretation of scores. The pseudo-independent variable was race, which was self-reported by participants in a demographic survey item.

Conducting the Analysis.

1. What test did they use, and why?

The authors used a one-way ANOVA to determine whether levels of outness would vary by racial group among queer and trans students of Color.

2. What are the assumptions of the test? Were they met in this case?

    a. Level of Measurement for the Dependent Variable is Interval or Ratio

    The dependent variable was outness, which was measured as averaged Likert-type data transformed into standard scores, so is measured at the interval level.

    b. Normality of the Dependent Variable

    The authors in this manuscript did not report information on the normality of the dependent variable. In the simulated dataset in the online resources, the data will be perfectly normal. However, in most cases, published papers will not report information on normality if this assumption was met. It is more typical to see a discuss of normality only when the data were not normally distributed. So, it is probably safe to infer from the fact the authors

did not report normality that the data were normally distributed. To test for normality, we would use skewness and kurtosis statistics as compared to their standard error, as described in Chapter 3.

c. Observations are Independent

The authors did not note any concerns with independence. There are potential nesting effects as the sample included students from multiple campuses, so there may be campus effects that were not captured in this research design. This may be difficult due to the dispersion of the sample among sites, as well.

d. Random Sampling and Assignment

The sample is not random. The authors used both purposive and snowball sampling, which included contacting leaders of LGBTQ groups, organizations, and campus liaisons to attempt to reach potential participants, in addition to their social media advertising.

The sample was not randomly assigned to groups, as the pseudo-independent variable was race. This involved the use of intact groups.

e. Homogeneity of variance

Using Levene's test, the assumption of homogeneity of variance was met ($F_{3, 382} = 0.248$, $p = .863$).

3. What was the result of that test?

There was a significant difference in outness between racial groups ($F_{3, 382} = 6.927$, $p < .001$).

4. What was the effect size, and how is it interpreted?

$$\omega^2 = \frac{SS_B - (df_B)(MS_w)}{SS_T + MS_w} = \frac{19.791 - (3)(0.952)}{383.584 + 0.952} = \frac{19.791 - 2.856}{384.536} = \frac{16.935}{384.536} = .044$$

About 4% of the variance in outness was explained by racial group ($\omega^2 = .044$).

5. What is the appropriate follow-up test, if any?

To determine how levels of outness varied among the four racial groups, we used Scheffe post-hoc tests.

6. What is the pattern of group differences?

Asian/Pacific Islander students had significantly lower levels of outness than Latinx ($p = .006$) or multiracial ($p = .008$) students. Black/African American students also had significantly lower levels of outness than Latinx ($p = .033$) or multiracial ($p = .047$) students. There was no significant difference between Asian/Pacific Islander and Black/African American students ($p = .913$), nor was there a significant difference between Latinx and multiracial students ($p = .947$).

Write up.

Results

We used a one-way ANOVA to determine whether levels of outness would vary by racial group among queer and trans students of Color. There was a significant difference in outness between racial groups ($F_{3, 382} = 6.927$, $p < .001$). About 4% of the variance in outness was explained by racial group ($\omega^2 = .044$). To determine how levels of outness varied among the four racial groups, we used Scheffe post-hoc tests. Asian/Pacific Islander students had significantly lower levels of outness than Latinx ($p = .006$) or multiracial ($p = .008$) students. Black/African American students also had significantly lower levels of outness than Latinx ($p = .033$) or multiracial ($p = .047$) students. There was no significant difference between Asian/Pacific Islander and Black/African American students ($p = .913$), nor was there a significant difference between Latinx and multiracial students ($p = .947$). See Table 1 for descriptive statistics by group. In the present sample, Latinx and multiracial students reported higher levels of outness than Asian/Pacific Islander and Black/African American students.

In APA style, tables can be placed in text or on a new page after the references page. For this example, we might make a table such as:

Table 1

Descriptive Statistics for Peer Support by Group

Group	*N*	*M*	*SD*
Asian/Pacific Islander	63	-.360	.980
Black/African American	78	-.240	1.030
Latinx	81	.220	.930
Multiracial	164	.140	.970
Total	386	-.002	.998

CASE STUDY 2: PRESERVICE TEACHERS' ATTITUDES TOWARDS DIVERSITY

Kumar, R., & Hamer, L. (2013). Preservice teachers' attitudes and beliefs toward Student Diversity and Proposed Instructional Practices: A Sequential Design Study. *Journal of Teacher Education*, *64*(2), 162–177. https://doi.org/10.1177/0022487112466899

In this article, the authors conducted a sequential design study to examine White preservice teacher's beliefs toward their culturally and economically diverse students. The authors argued that preservice teachers who are exposed to curricula that is predicated upon multicultural education and practices during their teacher education preparation program, are less biased and prejudiced towards their poor and minority students.

Research Questions.

The authors focused on these three research questions:

2. Is there a difference in attitudes and beliefs regarding stereotypes of race based on a student's level in the licensure program?
3. Is there a difference in attitudes and beliefs regarding stereotypes of socioeconomic status based on student's level in the licensure program?
4. Is a difference in attitudes and beliefs regarding discomfort with diversity based on student's level in the licensure program?

Hypotheses.

The authors hypothesized the following related to stereotypes of race:

H_0: There was no significant difference between preservice teachers' beliefs and attitudes regarding stereotypes of race based on their level in the program.

H_1: There was a significant difference between preservice teachers' beliefs and attitudes regarding stereotypes of race based on their level in the program

The authors hypothesized the following related to stereotypes of SES:

H_0: There was no significant difference between preservice teachers' beliefs and attitudes regarding stereotypes of SES based on their level in the program.

H_1: There was a significant difference between preservice teachers' beliefs and attitudes regarding stereotypes of SES based on their level in the program

The authors hypothesized the following related to discomfort with diversity:

H_0: There was no significant difference between preservice teachers' beliefs and attitudes regarding discomfort with diversity based on their level in the program.

H_1: There was a significant difference between preservice teachers' beliefs and attitudes regarding discomfort with diversity based on their level in the program

Variables Being Measured.

The authors conducted a sequential design, that involved cross-sectional and longitudinal data (a total of 4 data collection waves), and cluster analysis included White teachers ($n = 784$). The survey included: Beliefs about ethnic minority students (five-item scale); Beliefs about LSES students (four-item scale).

Conducting the Analysis.

7. What test did they use, and why?

The authors used a one-way ANOVA to determine if there were differences in preservice teachers' beliefs regarding stereotypes of race, stereotypes of SES, and discomfort with diversity based on their level in the licensure program.

8. What are the assumptions of the test? Were they met in this case?
    a. Level of Measurement for the Dependent Variable is Interval or Ratio

    The three dependent variables for this study were stereotypes of race, stereotypes of SES, and discomfort with diversity, all of which were interval scales.

    b. Normality of the Dependent Variable

    In most cases, when this assumption is met, the article will not report normality statistics. Because the authors did not report about normality, we must infer this assumption was met. However, normally researchers would evaluate this assumption before running the analysis (even if they do not write about it in the article). We would check for normality using skewness and kurtosis statistics.

    c. Observations are Independent

    The authors did not report any concerns with independence. There is no evidence of dependence as the surveys were administered and completed individually.

    d. Random Sampling and Assignment

    The sample appears to be a convenience sample because all of the participants are a particular set of students from a Midwestern University.

    There was no random assignment because they used intact groups based on level in program.

    e. Homogeneity of variance

The authors did not report whether or not the assumption for homogeneity of variance was met or not. Using Levene's test, we found that the assumption was not met for stereotypes of race ($F_{2,781} = 4.068$, $p = .017$), nor stereotypes of SES ($F_{2,781} = 5.167$, $p = .006$). However, the assumption was met for discomfort with diversity ($F_{2,781} = 1.061$, $p = .347$).

9. What was the result of that test?

There was a significant difference in preservice teachers' beliefs on stereotypes of race ($F_{2,781} = 3.574$, $p = .028$), stereotypes of SES ($F_{2,781} = 4.161$, $p = .016$), and discomfort with diversity ($F_{2,781} = 14.739$, $p < .001$) based on their level in the program.

10. What was the effect size, and how is it interpreted?

$$Race: \omega^2 = \frac{SS_B - (df_B)(MS_w)}{SS_T + MS_w} = \frac{3.398 - (2)(.475)}{374.628 + .475} = \frac{3.398 - .950}{375.103} = \frac{2.448}{375.103} = .007$$

About 1% of the variance in preservice teachers' beliefs on stereotypes of race was explained by their level in the program ($\omega^2 = .007$).

$$SES: \omega^2 = \frac{SS_B - (df_B)(MS_w)}{SS_T + MS_w} = \frac{3.987 - (2)(.479)}{378.216 + .479} = \frac{3.987 - .958}{378.695} = \frac{3.029}{378.695} = .008$$

About 1% of the variance in preservice teachers' beliefs on stereotypes of SES was explained by their level in the program ($\omega^2 = .008$).

$$Diversity: \omega^2 = \frac{SS_B - (df_B)(MS_w)}{SS_T + MS_w} = \frac{8.823 - (2)(.299)}{242.584 + .299} = \frac{8.823 - .598}{242.883} = \frac{8.225}{242.883} = .034$$

About 3% of the variance in preservice teachers' beliefs on discomfort with diversity was explained by their level in the program ($\omega^2 = .034$).

11. What is the appropriate follow-up test, if any?

As the omnibus test showed significance, the appropriate follow-up test was to conduct post-hoc tests, in which we used the Sheffe correction.

12. What is the pattern of group differences?

When examining preservice teachers' beliefs on stereotypes of race, there was a significant difference between students enrolled in the OS and STS courses ($p = .037$). There was no significant difference between students enrolled in OS and SDS courses ($p = .628$), nor between students enrolled in SDS and STS courses ($p = .159$). Similarly, when examining their beliefs on stereotypes of SES, there was a significant difference between students enrolled in the OS and STS courses ($p = .017$). There was no significant difference between students enrolled in OS and SDS courses ($p = .355$), nor between students enrolled in SDS and STS courses ($p = .222$). In regard to their discomfort with diversity, there was a significant difference between

students enrolled in the OS and STS courses ($p < .001$) and between students enrolled in the SDS and STS courses ($p < 0.001$). There was no significant difference between students enrolled in OS and SDS courses ($p = .478$). Those that were enrolled in the OS course had significantly higher ratings than those enrolled in STS courses, however the differences were slight.

Write up.

Results

The authors used a one-way ANOVA to determine if there were differences in preservice teachers' beliefs regarding stereotypes of race, stereotypes of SES, and discomfort with diversity based on their level in the licensure program. There was a significant difference in preservice teachers' beliefs on stereotypes of race ($F_{2,781} = 3.574$, $p = .028$), stereotypes of SES ($F_{2,781} = 4.161$, $p = .016$), and discomfort with diversity ($F_{2,781} = 14.739$, $p < .001$) based on their level in the program. About 1% of the variance in preservice teachers' beliefs on stereotypes of race was explained by their level in the program. Similarly, about 1% of the variance in preservice teachers' beliefs on stereotypes of SES was explained by their level in the program. About 3% of the variance in preservice teachers' discomfort with diversity was explained by their level in the program.

As the omnibus test showed significance, the appropriate follow-up test was to conduct post-hoc tests, in which we used the Sheffe correction. When examining preservice teachers' beliefs on stereotypes of race, there was a significant difference between students enrolled in the OS and STS courses ($p = .037$). There was no significant difference between students enrolled in OS and

SDS courses (p = .628), nor between students enrolled in SDS and STS courses (p = .159). Similarly, when examining their beliefs on stereotypes of SES, there was a significant difference between students enrolled in the OS and STS courses (p = .017). There was no significant difference between students enrolled in OS and SDS courses (p = .355), nor between students enrolled in SDS and STS courses (p = .222). In regard to their discomfort with diversity, there was a significant difference between students enrolled in the OS and STS courses (p < .001) and between students enrolled in the SDS and STS courses (p < 0.001). There was no significant difference between students enrolled in OS and SDS courses (p = .478). Those that were enrolled in the OS course had significantly higher ratings than those enrolled in STS courses, however the differences were slight.

FACTORIAL ANOVA CASE STUDIES

In this chapter, we will present several examples of published research that used the factorial ANOVA. For each sample, we encourage you to:

1) Use your library resources to find the original, published article. Read that article and look for how they use and talk about the factorial ANOVA.
2) Visit this book's online resources and download the datasets that accompany this chapter. Each dataset is simulated to reproduce the outcomes of the published research. (Note: the online datasets are not real human subjects data but have been simulated to match the characteristics of the published work.)
3) Follow along with each step of the analysis, comparing your own results with what we provide in this chapter. This will help cement your understanding of how to use the analysis.

CASE STUDY 1: INTERRACIAL FEEDBACK CONFLICTS

Harber, K. D., Reeves, S., Goman, J. L., Williams, C. H., Malin, J., & Pennebaker, J. W. (2018). The conflicted language of interracial feedback. *Journal of Educational Psychology, 111*(7), 1220-1242. https://doi.org/10.1037/edu0000326

In this manuscript, the authors evaluate the nature of feedback given by white evaluators when responding to a poorly written essay. Evaluators were told they were evaluating the essay of either a Black or a white student. Participants (who served as evaluators) were also randomly assigned to a self-image condition, which involved a social issues survey that had three versions: self-image affirmed, self-image neutral, and self-image threatened. In the larger context of the study, the authors found that white evaluators gave Black writers more positive feedback in tone but that was more lenient and less analytic in content. For this case study, we analyze one specific research question.

Research Questions.

Was the percentage of positively worded copyedit comments significantly different based on the interaction of the randomly assigned race of the essay writer (Black or white author description) and the self-image condition (affirming, neutral, or threatening)?

Hypotheses.

The authors hypothesized the following:

H_0: There was no significant difference in the percentage of positively worded copyedit comments based on the interaction of the randomly assigned race of the essay writer (Black or white author description) and the self-image

condition (affirming, neutral, or threatening). ($M_{\text{BlackXAffirmed}} = M_{BlackXNeutral} = M_{\text{BlackXThreatened}} = M_{\text{WhiteXAffirmed}} = M_{\text{WhiteXNeutral}} = M_{\text{WhiteXThreatened}}$).

H_1: There was no significant difference in the percentage of positively worded copyedit comments based on the interaction of the randomly assigned race of the essay writer (Black or white author description) and the self-image condition (affirming, neutral, or threatening). ($M_{\text{BlackXAffirmed}} \neq M_{BlackXNeutral} \neq M_{\text{BlackXThreatened}} \neq M_{\text{WhiteXAffirmed}} \neq M_{\text{WhiteXNeutral}} \neq M_{\text{WhiteXThreatened}}$).

Variables Being Measured.

The dependent variable for this research question was the percentage of positively worded copyedit comments. The copyedit comments left by participants evaluating the essay were independently coded by two individuals on several characteristics, including whether the comments were positively worded. The two coders' scores were averaged to create the percentage of positive comments. The two independent variables were the essay author's race and the self-image condition. For essay author's race, all participants evaluated the same essay, but were randomly assigned to be told the author was either Black or white. The self-image condition was also randomly assigned. Participants were randomly assigned one of three versions of the Social Issues Survey designed to either affirm, threaten, or be neutral toward their self-image.

Conducting the Analysis.

1. What test did they use, and why?

The authors used a factorial ANOVA to determine if percentage of positively worded copyedit comments were significantly different based on the interaction of the randomly assigned race of the essay writer (Black or white author description) and the self-image condition (affirming, neutral, or threatening).

2. What are the assumptions of the test? Were they met in this case?

    a. Level of Measurement for the Dependent Variable is Interval or Ratio

    The dependent variable is calculated as a percentage, so the data are ratio.

    b. Normality of the Dependent Variable

    The authors did not report data on normality, which is typical in journal articles if the assumption of normality was met. In practice, we would test for normality as a preliminary step in the analysis, using skewness and kurtosis statistics, even if we did not ultimately include that information in the published article.

    c. Observations are Independent

    The authors note no factors that threaten independence.

    d. Random Sampling and Assignment

The sample was not random. It was comprised of undergraduate psychology students and was a convenience sample. The participants were, however, randomly assigned to groups on both independent variables.

e. Homogeneity of variance

The assumption of homogeneity of variance was not met ($F_{5, 95} = 16.283$, $p < .001$). All cell sizes were relatively equal, with the largest cell having $n = 19$ and the smallest $n = 15$. In the published study, the authors did not use a correction for heterogeneity of variance. The failure of this assumption likely attenuates the omnibus test value.

3. What was the result of that test?

There was a significant difference in the percentage of positive feedback based on the interaction ($F_{2, 95} = 4.955$, $p = .009$).[2]

4. What was the effect size, and how is it interpreted?

$$\omega^2 = \frac{SS_E - (df_E)(MS_w)}{SS_T + MS_w} = \frac{172.291 - (2)(17.386)}{1906.481 + 17.386} = \frac{172.291 - 34.772}{1923.867} = \frac{137.519}{1923.867} = .071$$

About 7% of the variance in the percentage of positive copyedit comments was explained by the interaction of race and self-image groups ($\omega^2 = .071$)

5. What is the appropriate follow-up analysis?

To explore the disordinal interaction and determine how cells differed from one another, we used simple effects analysis.

6. What is the result of the follow-up analysis?

Among those from the self-image affirmed group, there was no difference based on the author's race ($F_{1, 95} = .186$, $p = .667$). Among the self-image neutral group, those rating the white author gave significantly more positive feedback ($F_{1, 95} = 4.059$, $p = .047$). Finally, among the self-image threatened group, those rating the Black author gave significantly more positive feedback ($F_{1, 95} = 5.666$, $p = .019$).

7. What is the pattern of group differences?

Among the present sample of White undergraduate students, participants gave more positive feedback to white authors in a self-image neutral condition, but more positive feedback to Black authors in a self-image threat condition.

[2] Please note that the values from the simulated data provided in the online course resources differ slightly from the authors' calculations. This is an artifact of the simulation process and the authors results are not incorrect or in doubt.

Write up.

Results

We used a factorial ANOVA to determine if percentage of positively worded copyedit comments were significantly different based on the interaction of the randomly assigned race of the essay writer (Black or white author description) and the self-image condition (affirming, neutral, or threatening). The assumption of homogeneity of variance was not met ($F_{5, 95} = 16.283$, $p < .001$). There was a significant difference in the percentage of positive feedback based on the interaction ($F_{2, 95} = 4.955$, $p = .009$). To explore the disordinal interaction and determine how cells differed from one another, we used simple effects analysis. Among those from the self-image affirmed group, there was no difference based on the author's race ($F_{1, 95} = .186$, $p = .667$). Among the self-image neutral group, those rating the white author gave significantly more positive feedback ($F_{1, 95} = 4.059$, $p = .047$). Finally, among the self-image threatened group, those rating the Black author gave significantly more positive feedback ($F_{1, 95} = 5.666$, $p = .019$). Among the present sample of White undergraduate students, participants gave more positive feedback to white authors in a self-image neutral condition, but more positive feedback to Black authors in a self-image threat condition.

Notice that, because the interaction was significant, all of our interpretive attention is on the interaction. In fact, we have not interpreted the main effects at all. We would also likely include tables and a figure. These could be placed after the references page on a new page, or within the text.

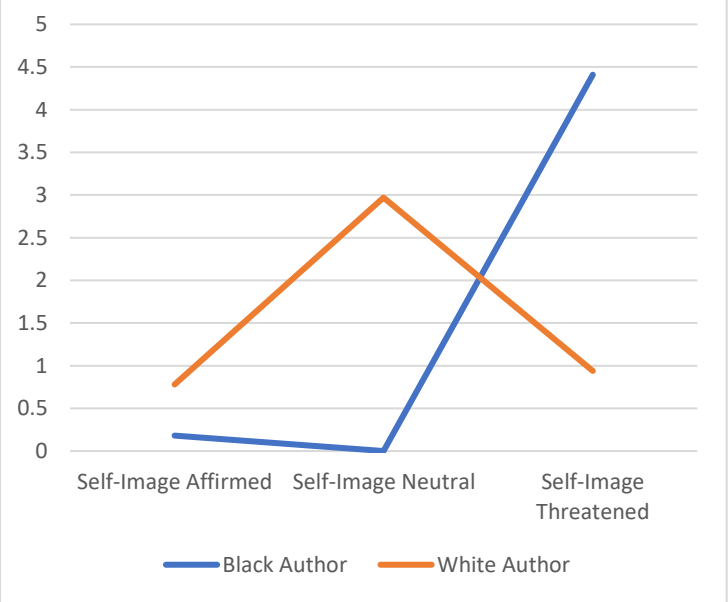
Table 1

Descriptive Statistics for Seriousness Ratings

Author Race	Self-Image Condition	*M*	*SD*	*N*
Black	Affirming	.180	.790	17
	Neutral	.000	.000	16
	Threatening	4.410	6.950	18
	Total	1.616	4.576	51
White	Affirming	.780	1.960	19
	Neutral	2.970	6.150	16
	Threatening	.940	3.640	15
	Total	1.529	4.216	50
Total	Affirming	.497	1.520	36
	Neutral	1.485	4.536	32
	Threatening	2.833	5.877	33
	Total	1.573	4.380	101

Figure 1.

Plot of Cell Means



CASE STUDY 2: STEREOTYPE THREAT AND READING PERFORMANCE

Wasserberg, M. J. (2014). Stereotype threat on African American children in an urban elementary school. *Journal of Experimental Education, 82*(4), 502-517. https://doi.org/10.1080/00220973.2013.876224

In this manuscript, the author tests how stereotype threat might influence Black students' performance on reading ability tests. Stereotype threat refers to a phenomenon where members of a group that is subject to negative stereotypes experience anxiety or fear that their behavior will confirm the negative stereotype. Researchers have demonstrated that stereotype threat can reduce performance, likely due to the anxiety it induces. In this study, the author sought to test differences between children who were aware or unaware of stereotypes and between testing with a diagnostic or nondiagnostic purpose.

Research Questions.

Did reading test performance significantly differ based on the interaction of stereotype awareness (aware versus unaware) and whether or not the test was presented as a diagnostic test?

Hypotheses.

The authors hypothesized the following:

H_0: There was no significant difference in reading test scores based on the interaction of stereotype awareness (aware versus unaware) and how the test was described (diagnostic versus nondiagnostic). ($M_{\text{AwareXDiagnostic}} = M_{AwareXNondiagnostic} = M_{\text{UnawareXDiagnostic}} = M_{\text{UnawareXNondiagnostic}}$).

H_1: There was a significant difference in reading test scores based on the interaction of stereotype awareness (aware versus unaware) and how the test was described (diagnostic versus nondiagnostic). ($M_{\text{AwareXDiagnostic}} \neq M_{AwareXNondiagnostic} \neq M_{\text{UnawareXDiagnostic}} \neq M_{\text{UnawareXNondiagnostic}}$).

Variables Being Measured.

The dependent variable for this research question was performance on a standardized reading comprehension test. The test involved reading a passage, followed by 13 multiple-choice items, then a second passage followed by another 8 multiple-choice questions. The first independent variable was stereotype awareness. The researcher measured whether participants were stereotype aware or unaware using a vignette and follow-up questions. The vignette presented a fictional world in which stereotypes were present. Participants then answered the extent to which that fictional world was like the real world. Based on their answers to that question, students were classified into the two stereotype awareness groups. The second independent variable was the nature of the test presentation (diagnostic or

nondiagnostic). Participants were randomly assigned to groups and told either that the test was being used to determine their ability level to find out how well they would do on state testing (diagnostic condition), or that the test was a problem-solving activity only (nondiagnostic condition). All participants were Black children in the 3rd, 4th, or 5th grades.

Conducting the Analysis.

1. What test did they use, and why?

The author used a factorial ANOVA to determine if test performance significantly differed based on the interaction of stereotype awareness (aware versus unaware) and whether or not the test was presented as a diagnostic test.

2. What are the assumptions of the test? Were they met in this case?
    a. Level of Measurement for the Dependent Variable is Interval or Ratio

    The dependent variable is calculated as a percentage, so the data are ratio.

    b. Normality of the Dependent Variable

    The author did not report data on normality, which is typical in journal articles if the assumption of normality was met. In practice, we would test for normality as a preliminary step in the analysis, using skewness and kurtosis statistics, even if we did not ultimately include that information in the published article.

    c. Observations are Independent

    The author noted no factors that threaten independence.

    d. Random Sampling and Assignment

    The sample was not random. All participants came from a single urban elementary school, so it is likely a convenience sample. Participants were not randomly assigned to stereotype awareness groups and were assigned to groups based on their answers to the stereotype vignette. However, participants were randomly assigned to the diagnostic or nondiagnostic test condition.

    e. Homogeneity of variance

    The assumption of homogeneity of variance was met ($F_{3, 139} = 2.083, p = .104$).
3. What was the result of that test?

There was a significant difference in reading comprehension test scores based on the interaction ($F_{1, 169} = 9.603, p = .002$).[3]

[3] Please note that the values from the simulated data provided in the online course resources differ slightly from the authors' calculations. This is an artifact of the simulation process and the authors' results are not incorrect or in doubt.

4. What was the effect size, and how is it interpreted?

$$\omega^2 = \frac{SS_E - (df_E)(MS_w)}{SS_T + MS_w} = \frac{3032.555 - (1)(315.805)}{63712.596 + 315.805} = \frac{3032.555 - 315.805}{64028.401}$$
$$= \frac{2716.75}{64028.401} = .042$$

About 4% of the variance in reading comprehension test scores was explained by the interaction of stereotype awareness and test condition ($\omega^2 = .042$)

5. What is the appropriate follow-up analysis?

To explore the disordinal interaction and determine how cells differed from one another, we used simple effects analysis.

6. What is the result of the follow-up analysis?

Among stereotype aware participants, those in the nondiagnostic test condition scored significantly higher than those in the diagnostic condition ($F_{1, 169} = 15.393$, $p < .001$). There was no significant difference between diagnostic and nondiagnostic conditions among stereotype unaware participants ($F_{1, 169} = 0.106$, $p = .745$).

7. What is the pattern of group differences?

Among the present sample of Black elementary school students, those who were aware of stereotypes performed better on a reading comprehension test when it was presented in nondiagnostic terms when compared to those to whom the test was presented as diagnostic.

Write up.

Results

We used a factorial ANOVA to determine if test performance significantly differed based on the interaction of stereotype awareness (aware versus unaware) and whether or not the test was presented as a diagnostic test. There was a significant difference in reading comprehension test scores based on the interaction ($F_{1, 169} = 9.603$, $p = .002$). About 4% of the variance in reading comprehension test scores was explained by the interaction of stereotype awareness and test condition ($\omega^2 = .042$). To explore the disordinal interaction and determine how cells differed from one another, we used simple effects analysis. Among stereotype aware participants, those in the nondiagnostic test condition

scored significantly higher than those in the diagnostic condition ($F_{1, 169} = 15.393$, $p < .001$). There was no significant difference between diagnostic and nondiagnostic conditions among stereotype unaware participants ($F_{1, 169} = 0.106$, $p = .745$). Among the present sample of Black elementary school students, those who were aware of stereotypes performed better on a reading comprehension test when it was presented in nondiagnostic terms when compared to those to whom the test was presented as diagnostic.

Notice that, because the interaction was significant, all of our interpretive attention is on the interaction. In fact, we have not interpreted the main effects at all. We would also likely include tables and a figure. These could be placed after the references page on a new page, or within the text.

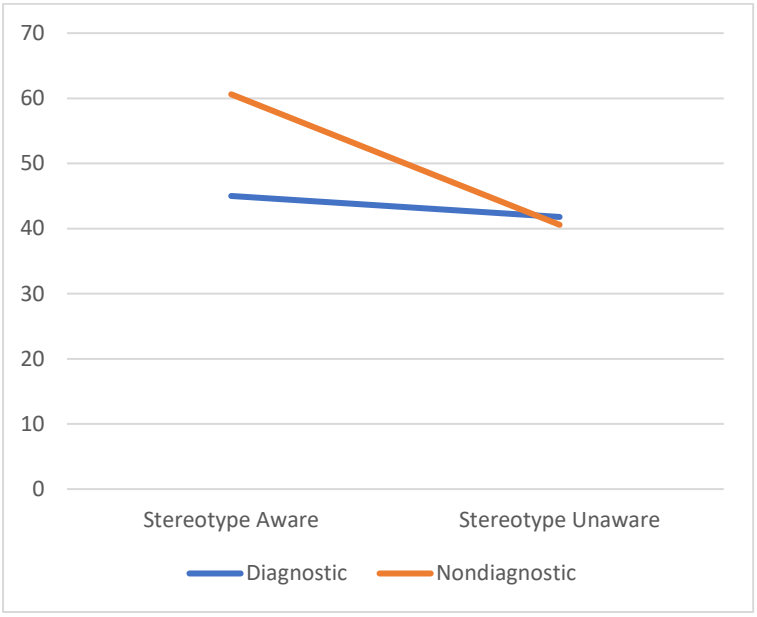
Table 1

Descriptive Statistics for Reading Performance

Stereotype Awareness	Test Condition	*M*	*SD*	*N*
Aware	Diagnostic	45.000	18.010	42
	Nondiagnostic	60.610	21.600	38
	Total	52.410	21.170	80
Unaware	Diagnostic	41.780	17.150	45
	Nondiagnostic	40.580	14.490	48
	Total	41.160	15.760	93
Total	Diagnostic	43.330	17.540	87
	Nondiagnostic	49.43	20.470	86
	Total	46.360	19.250	173

Figure 1.

Plot of Cell Means



PAIRED SAMPLES *t*-TEST CASE STUDIES

In this chapter, we will present several examples of published research that used the paired samples *t*-test. We should note that, in these examples, the simulated data provided in the online resources will not produce the exact result of the published study. However, they will reproduce the essence of the finding – so don't be surprised to look up the published study and see somewhat different results.[4] For each sample, we encourage you to:

1) Use your library resources to find the original, published article. Read that article and look for how they use and talk about the *t*-test.
2) Visit this book's online resources and download the datasets that accompany this chapter. Each dataset is simulated to reproduce the outcomes of the published research. (Note: the online datasets are not real human subjects data but have been simulated to match the characteristics of the published work.)
3) Follow along with each step of the analysis, comparing your own results with what we provide in this chapter. This will help cement your understanding of how to use the analysis.

CASE STUDY 1: PRE-SERVICE FOREIGN LANGUAGE TEACHER ATTITUDES ABOUT OPPRESSION

McGowan, S. L., & Kern, A. L. (2014). Pre-service foreign language teachers' attitudes of privilege and oppression. *Journal of Education and Training Studies, 2*(1), 31-43. https://doi.org/10.11114/jets.v2i1.188

[4] We simulate the data for the online resources by simulating data with a certain mean and standard deviation. That works perfectly for the between-subjects designs, which really measure only mean differences. But, for within-subjects designs like the paired samples *t*-test, this turns out fairly differently. The paired samples *t*-test uses the mean of differences per case, rather than the mean difference overall. Because we do not know the mean difference per case from published work, we cannot simulate data that perfectly reproduce those results. However, the overall mean difference and the direction of the result will be the same as the published study. In most cases, this results in a smaller effect size for the simulated data in the online resources than for the actual published study. The published results are not in doubt, but we cannot perfectly reproduce them in our simulated data.

In this study, the researchers used a pre-test/post-test design to determine whether a course that included intentional discussion of white privilege, racism, and oppression would be associated with changes in pre-service teachers' attitudes about privilege and oppression. In the full study, they measure this change overall, in attitudes about white privilege, in attitudes about heterosexism, and in attitudes about Christian privilege. In this case study, we focus on the overall score.

Research Questions.

Did pre-service teachers' attitudes about privilege and oppression improve significantly from before the foreign language course to after it?

Hypotheses.

H_0: There was no significant improvement in attitudes about privilege from pre-test to post-test. ($M_{pre} \geq M_{post}$)

H_1: There was a significant improvement in attitudes about privilege from pre-test to post-test. ($M_{pre} < M_{post}$)

Variables Being Measured.

The authors used a 39-item Likert-type scale called the Privilege and Oppression Inventory to measure attitudes about privilege and oppression. The independent variable was time, which involved administering the survey before and after the course to test for differences.

Conducting the Analysis.

8. What test did they use, and why?

The authors used a paired samples *t*-test to determine if attitudes about privilege and oppression would improve significantly from before to after a course that included intentional discussion of privilege, oppression, and racism. (Note that, because the hypothesis is directional [it hypothesizes increases in scores from pre- to post-test] that this is a one-tailed test).

9. What are the assumptions of the test? Were they met in this case?

    a. Level of Measurement for the Dependent Variable is Interval or Ratio

    The overall attitude about privilege and oppression was calculated as an average of the 39 Likert-type items, so those data can be considered interval.

    b. Normality of the Dependent Variable

    The authors do not discuss normality in the published paper. This is typical when the data are normally distributed. Ideally, we would test for normality prior to running the analysis by using skewness and kurtosis statistics, even if those will not ultimately be reported in the manuscript. As a reminder, the

data on the course online resources will be normally distributed because of how they were simulated.

c. Observations are Independent

The authors note no potential nesting factors or other factors that might cause dependence. However, because this is a pre-test/post-test design, there is a possibility of order effects or practice effects.

d. Random Sampling and Assignment

Participants were not randomly sampled and appear to be a convenience sample of university students. They were also not randomly assigned to order of administration because the design was longitudinal (pre-test vs. post-test) so counterbalancing the order of administration was not possible.

10. What was the result of that test?

There was a significant improvement in attitudes about privilege and oppression (t_{18} = -2.174, p = .022).[5]

11. What was the effect size, and how is it interpreted?

$$\omega^2 = \frac{t^2 - 1}{t^2 + n - 1} = \frac{-2.174^2 - 1}{-2.174^2 + 19 - 1} = \frac{4.726 - 1}{4.726 + 18} = \frac{3.726}{22.726} = .164$$

About 16% of the variance in attitudes about privilege and oppression was explained by the change from before the course to after the course (ω^2 = .164).

12. What is the pattern of group differences?

Attitudes about privilege and oppression had higher scores after the course (M = 4.860, SD = .560) than before the course (M = 4.660, SD = .580).

Write up.

Results

We used a paired samples *t*-test to determine if attitudes about privilege and oppression would improve significantly from before to after a course that included intentional discussion of privilege, oppression, and racism. There was a

[5] As a reminder, this value will not match the published study exactly. As a second note about this value, the author reports a positive *t*-test value, so may have dropped the sign and reported the absolute value or set the comparison up as post- vs. pre-test instead of pre-test vs. post-test. In addition, because this is a directional hypothesis (one-tailed test), we would divide the probability values in half to get the one-tailed probabilities

significant improvement in attitudes about privilege and oppression (t_{18} = -2.174, p = .022). About 16% of the variance in attitudes about privilege and oppression was explained by the change from before the course to after the course (ω^2 = .164). Attitudes about privilege and oppression had higher scores after the course (M = 4.860, SD = .560) than before the course (M = 4.660, SD = .580).

CASE STUDY 2: CULTURALLY RESPONSIVE TEACHING IN SOCIAL STUDIES

Fitchett, P. G., Starker, T. V., & Salyers, B. (2012). Examining culturally responsive teaching self-efficacy in a preservice social studies education course. *Urban Education, 47*(3), 585-611. https://doi.org/10.1177/0042085912436568

The researchers in this study examined pre-service social studies teachers' confidence or self-efficacy in implementing culturally responsive teaching in their courses. They used a pre-test/post-test design to evaluate how participants' confidence changed from before to after they were taught using a model of culturally responsive teaching called the 3Rs model. They measured several different aspects of attitude and confidence, but this case study will focus on their examination of culturally responsive teaching confidence.

Research Questions.

Did pre-service teachers' confidence in implementing culturally responsive teaching in a social studies course improve significantly from before the 3Rs instruction to after it?

Hypotheses.

H_0: There was no significant improvement in pre-service teachers' confidence in implementing culturally responsive teaching in a social studies course from pre-test to post-test. ($M_{pre} \geq M_{post}$)

H_1: There was a significant improvement in pre-service teachers' confidence in implementing culturally responsive teaching in a social studies course from pre-test to post-test. ($M_{pre} < M_{post}$)

Variables Being Measured.

The authors used a 40-item Likert-type scale called the Culturally Responsive Teaching Self-Efficacy Scale to measure confidence in implementing culturally responsive teaching practices. The independent variable was time, which involved administering the survey before and after instruction in the 3Rs model of culturally responsive teaching.

Conducting the Analysis.

1. What test did they use, and why?

The authors used a paired samples *t*-test to determine if pre-service teachers' confidence in implementing culturally responsive teaching in a social studies course improved significantly from before the 3Rs instruction to after it. (Note that, because the hypothesis is directional [it hypothesizes increases in scores from pre- to post-test] that this is a one-tailed test).

2. What are the assumptions of the test? Were they met in this case?
    a. Level of Measurement for the Dependent Variable is Interval or Ratio

The confidence in implementing culturally responsive teaching practices was calculated as a sum of the 40 Likert-type items, so those data can be considered interval.

b. Normality of the Dependent Variable

The authors do not discuss normality in the published paper. This is typical when the data are normally distributed. Ideally, we would test for normality prior to running the analysis by using skewness and kurtosis statistics, even if those will not ultimately be reported in the manuscript. As a reminder, the data on the course online resources will be normally distributed because of how they were simulated.

c. Observations are Independent

The authors note no potential nesting factors or other factors that might cause dependence. However, because this is a pre-test/post-test design, there is a possibility of order effects or practice effects.

d. Random Sampling and Assignment

Participants were not randomly sampled and appear to be a convenience sample of university students. They were also not randomly assigned to order of administration because the design was longitudinal (pre-test vs. post-test) so counterbalancing the order of administration was not possible.

3. What was the result of that test?

There was a significant improvement in confidence in implementing culturally responsive social studies teaching practices (t_{19} = -3.479, p = .003).[6]

4. What was the effect size, and how is it interpreted?

$$\omega^2 = \frac{t^2 - 1}{t^2 + n - 1} = \frac{-3.479^2 - 1}{-3.479^2 + 20 - 1} = \frac{12.103 - 1}{12.103 + 19} = \frac{11.103}{31.103} = .357$$

About 36% of the variance confidence scores was explained by the change from before the 3Rs instruction to after that instruction (ω^2 = .357).

5. What is the pattern of group differences?

Participant's confidence in implementing culturally responsive social studies teaching practices was higher after the 3Rs instruction (M = 425.880, SD = 71.270) than it was before the instruction (M = 380.690, SD = 71.270).

[6] As a reminder, this value will not match the published study exactly. As a second note about this value, the author reports a positive t-test value, so may have dropped the sign and reported the absolute value or set the comparison up as post- vs. pre-test instead of pre-test vs. post-test. In addition, because this is a directional hypothesis (one-tailed test), we would divide the probability values in half to get the one-tailed probabilities

Write up.

Results

We used a paired samples *t*-test to determine if pre-service teachers' confidence in implementing culturally responsive teaching in a social studies course improved significantly from before the 3Rs instruction to after it. There was a significant improvement in confidence in implementing culturally responsive social studies teaching practices ($t_{19} = -3.479$, $p = .003$). About 36% of the variance confidence scores was explained by the change from before the 3Rs instruction to after that instruction ($\omega^2 = .357$). Participant's confidence in implementing culturally responsive social studies teaching practices was higher after the 3Rs instruction ($M = 425.880$, $SD = 71.270$) than it was before the instruction ($M = 380.690$, $SD = 71.270$).

WITHIN-SUBJECTS ANOVA CASE STUDIES

In the previous chapter, we explored the within-subjects ANOVA using a made-up example and some fabricated data. In this chapter, we will present several examples of published research that used the within-subjects ANOVA. For each sample, we encourage you to:

1) Use your library resources to find the original, published article. Read that article and look for how they use and talk about the within-subjects ANOVA.
2) Visit this book's online resources and download the datasets that accompany this chapter. Each dataset is simulated to reproduce the outcomes of the published research. (Note: the online datasets are not real human subjects data but have been simulated to match the characteristics of the published work.)
3) Follow along with each step of the analysis, comparing your own results with what we provide in this chapter. This will help cement your understanding of how to use the analysis.

CASE STUDY 1: DREAMER ALLY COMPETENCY

Cisneros, J., & Cadenas, G. (2017). DREAMer-ally competency and self-efficacy: Developing higher education staff and measuring lasting outcomes. *Journal of Student Affairs Research and Practice, 54*(2), 189-203. https://doi.org/10.1080/19496591.2017.1289098

In this article, the researchers investigated the learning outcomes of a program designed to improve higher education staff competency and self-efficacy for serving undocumented students (who are also referred to as DREAMers). They tracked participants in the training program (called DREAMzone) for eight months to evaluate whether the gains associated with the training would be sustained over time.

Research Questions.

The authors asked several research questions, but in this case study, we will focus on one: Did competency for working with undocumented students change from before the DREAMzone to three time points after the training (immediately after, two months later, and eight months later)?

Hypotheses.

The authors hypothesized the following related to competency to work with undocumented students:

H_0: There was no significant difference in competency scores between the pre-test, post-test, two-month follow-up, and eight-month follow-up. ($M_{pre} = M_{post} = M_{2\text{-month-follow-up}} = M_{8\text{-month-follow-up}}$)

H_1: There was a significant difference in competency scores between the pre-test, post-test, two-month follow-up, and eight-month follow-up. ($M_{pre} \neq M_{post} \neq M_{2\text{-month-follow-up}} \neq M_{8\text{-month-follow-up}}$).

Notice that, although the authors theorized that scores would improve at post-test and at the follow-up, the formal hypothesis does not specify a direction. The ANOVA design doesn't allow any specification of directionality in the omnibus test.

Variables Being Measured.

The authors measured competency in serving undocumented students using the DREAMer-Ally Competency scale. It was a six-item scale using 4-point Likert-type items and was specifically designed for this study. The authors report acceptable score reliability estimates at all four time points, ranging from $\alpha = .775$ to $\alpha = .867$. The independent variable was time, which included a pre-test, a post-test immediately following the training, a two-month follow-up and a four-month follow-up. The authors note substantial attrition in the sample size across the follow-up tests, which is not unusual for longitudinal data. However, that loss of participants over time should be taken into consideration when describing the results and their meaningfulness.

Conducting the Analysis.

6. What test did they use, and why?

The authors used the within-subjects ANOVA to determine if there was a significant difference in competency scores between the pre-test, post-test, two-month follow-up, and eight-month follow-up.

7. What are the assumptions of the test? Were they met in this case?

    a. Level of Measurement for the Dependent Variable is Interval or Ratio

    The dependent variable is measured using averaged Likert-type data, so is interval.

    b. Normality of the Dependent Variable

    The authors provide skewness and kurtosis statistics in the manuscript, finding that the data met the assumption of normality in all cases. (As a reminder, the simulated dataset available on the website will be perfectly normal due to the process by which we simulate the data.)

    c. Observations are Independent

    The authors did not note any issues of dependence in the data, or any nesting factors.

d. Random Sampling and Assignment

This is a convenience sample – all participants were from a single, large, public research university. The participants were not randomly assigned to order of administration (no counterbalancing) because the design was longitudinal.

e. Sphericity

The assumption of sphericity was met ($W_3 = .860$, $p = .159$).

8. What was the result of that test?

There was a significant difference in competency to serve undocumented students across the four tests ($F_{3, 162} = 53.306$, $p < .001$).[7]

9. What was the effect size, and how is it interpreted?

About 50% of the variance in competency scores was explained by the change between pre-test, post-test, two-month follow-up, and eight-month follow-up test ($\eta^2 = .497$).[8]

10. What is the appropriate follow-up analysis?

To determine how scores varied across the pre-test, post-test, and follow-up test, we used Tukey pairwise comparisons.

11. What is the result of the follow-up analysis?

There was a significant difference between pre-test and post-test ($p < .001$), two-month follow-up ($p < .001$), and eight-month follow-up ($p < .001$). There was no significant difference between post-test and the two-month follow-up ($p = .804$) or the eight-month follow-up ($p = .199$). Finally, there was no significant difference between the two-month and eight-month follow-up ($p = .703$).[9]

12. What is the pattern of group differences?

Compared to pre-test competency scores ($M = 15.345$, $SD = 3.092$), scores were higher at post-test ($M = 21.181$, $SD = 2.427$), the two-month follow-up test ($M = 20.709$, $SD = 2.362$), and the

---

[7] This value will not match the published results exactly. Because of the process used to simulate data for the online course page, it will not exactly match for within-subjects designs. The authors' published results are not in question here, but our simulated outcomes are slightly different.

[8] Again, note that this value will not match published values exactly. That is an artifact of the way that we have simulated data for the online course resources to allow students to practice the analysis, not a commentary on the published results.

[9] This, again, will not match the published results exactly due to the simulation process for the practice data in the online resources. Notably, the authors found a difference between post-test and eight-month follow-up that was small but statistically significant. They used a slightly different follow-up test, and again the simulation process does not perfectly reproduce the results for the within-subjects designs.

eight-month follow-up test (M = 20.145, SD = 2.606). These results appear to demonstrate gains in competency to work with undocumented students following the DREAMzone training that were relatively stable across time.

Write up.

Results

We used the within-subjects ANOVA to determine if there was a significant difference in competency scores between the pre-test, post-test, two-month follow-up, and eight-month follow-up. There was a significant difference in competency to serve undocumented students across the four tests ($F_{3, 162}$ = 53.306, $p < .001$). About 50% of the variance in competency scores was explained by the change between pre-test, post-test, two-month follow-up, and eight-month follow-up test (η^2 = .497). There was a significant difference between pre-test and post-test ($p < .001$), two-month follow-up ($p < .001$), and eight-month follow-up ($p < .001$). There was no significant difference between post-test and the two-month follow-up (p = .804) or the eight-month follow-up (p = .199). Finally, there was no significant difference between the two-month and eight-month follow-up (p = .703). Compared to pre-test competency scores (M = 15.345, SD = 3.092), scores were higher at post-test (M = 21.181, SD = 2.427), the two-month follow-up test (M = 20.709, SD = 2.362), and the eight-month follow-up test (M = 20.145, SD = 2.606). These results appear to demonstrate gains in competency to work with undocumented students following the DREAMzone training that were relatively stable across time.

CASE STUDY 2: AWARENESS OF RACIAL MICROAGGRESSIONS

Banks, B. M., Adams, D. F., Williams, C., & Piña, D. (2020). Preliminary investigation of efforts to improve awareness of racial microaggressions on campus. *Journal of Underrepresented and Minority Progress, 4*(1), 20-43. https://doi.org/10.32674/jump.v4i1.1763

In this manuscript, the authors measured responses to a bystander intervention workshop that was designed to teach participants how to combat racial microaggressions. They measured participants before the workshop, afterward, and then at three- and seven-week follow-ups. The authors intended the study as a preliminary investigation into whether a workshop might provide some long-term benefit in the awareness of racial microaggressions. They also hoped that increased awareness might lead to more bystander intervention to reduce microaggressions.

Research Questions.

The authors asked several research questions, but in this case study, we will focus on one: Did participants' knowledge about racial microaggressions change over time (before the workshop, immediately afterward, three weeks later, and seven weeks later)?

Hypotheses.

The authors hypothesized the following related to knowledge about racial microaggressions:

H_0: There was no significant difference in knowledge about racial microaggressions scores between the pre-test, post-test, three-week follow-up, and seven-week follow-up. ($M_{pre} = M_{post} = M_{3\text{-week-follow-up}} = M_{7\text{-week-follow-up}}$).

H_1: There was no significant difference in knowledge about racial microaggressions scores between the pre-test, post-test, three-week follow-up, and seven-week follow-up. ($M_{pre} \neq M_{post} \neq M_{3\text{-week-follow-up}} \neq M_{7\text{-week-follow-up}}$).

Notice that, although the authors theorized that scores would improve at post-test and remain stable at the follow-up tests, the formal hypothesis does not specify a direction. The ANOVA design doesn't allow any specification of directionality in the omnibus test, though those hypotheses could be specified as a priori comparisons.

Variables Being Measured.

The authors measured knowledge of racial microaggressions using a scale designed for the purposes of this study. The knowledge items consisted of six multiple-choice questions that were scored as either correct or incorrect. Scores were calculated as the number of correct items. The independent variable was time, which included a pre-test, a post-test immediately following the workshop, a three-week follow-up and a seven-week follow-up.

Conducting the Analysis.

1. What test did they use, and why?

The authors used the within-subjects ANOVA to determine if there was a significant difference in knowledge about racial microaggressions scores between the pre-test, post-test, three-week follow-up, and seven-week follow-up.

2. What are the assumptions of the test? Were they met in this case?

    a. Level of Measurement for the Dependent Variable is Interval or Ratio

    The dependent variable is measured as the number of correct answers on a multiple-choice quiz, so is interval.

    b. Normality of the Dependent Variable

    The authors do not discuss normality in the published paper. This is typical when the data are normally distributed. Ideally, we would test for normality prior to running the analysis by using skewness and kurtosis statistics, even if those will not ultimately be reported in the manuscript. As a reminder, the data on the course online resources will be normally distributed because of how they were simulated.

    c. Observations are Independent

    The authors did not note any issues of dependence in the data, or any nesting factors.

    d. Random Sampling and Assignment

    This is a convenience sample – all participants were from a single university. They were recruited via emails to a list of students who expressed interest in participating in research studies on campus. The participants were not randomly assigned to order of administration (no counterbalancing) because the design was longitudinal.

    e. Sphericity

    The assumption of sphericity was met ($W_3 = .822$, $p = .096$).

3. What was the result of that test?

There was a significant difference in knowledge about racial microaggressions across the four tests ($F_{3, 147} = 51.386$, $p < .001$).[10]

4. What was the effect size, and how is it interpreted?

[10] This value will not match the published results exactly. Because of the process used to simulate data for the online course page, it will not exactly match for within-subjects designs. The authors' published results are not in question here, but our simulated outcomes are slightly different.

About 51% of the variance in knowledge about racial microaggression scores was explained by the change between pre-test, post-test, three-week follow-up, and seven-week follow-up test ($\eta^2 = .512$).[11]

5. What is the appropriate follow-up analysis?

To determine how scores varied across the pre-test, post-test, and follow-up test, we used Scheffe pairwise comparisons.

6. What is the result of the follow-up analysis?

There was a significant difference between pre-test and post-test ($p < .001$), three-week follow-up ($p < .001$), and seven-week follow-up ($p < .001$). There was no significant difference between post-test and the three-week follow-up ($p = .856$) or the seven-week follow-up ($p = .970$). Finally, there was no significant difference between the three-week and seven-week follow-up ($p = .986$).[12]

7. What is the pattern of group differences?

Compared to pre-test knowledge scores ($M = 2.490$, $SD = 1.510$), scores were higher at post-test ($M = 5.580$, $SD = 1.610$), the three-week follow-up test ($M = 5.840$, $SD = 1.610$), and the seven-week follow-up test ($M = 5.720$, $SD = 1.730$). These results suggest increases in knowledge about racial microaggressions following the workshop that remained relatively stable at three- and seven-week follow-ups.

Write up.

Results

We used the within-subjects ANOVA to determine if there was a significant difference in knowledge about racial microaggressions scores between the pre-test, post-test, three-week follow-up, and seven-week follow-up. There was a significant difference in knowledge about racial microaggressions across the four tests ($F_{3, 147} = 51.386$, $p < .001$). About 51% of the variance in knowledge about racial microaggression scores was explained by the change between pre-test, post-test, three-week follow-up, and seven-week follow-up test ($\eta^2 = .512$). To determine how scores varied across the pre-test, post-test, and follow-up test, we used Scheffe pairwise comparisons. There was a significant difference between pre-test and post-test ($p < .001$), three-week follow-up ($p < .001$), and seven-week follow-up ($p < .001$). There was no significant difference between post-test and the three-week follow-up ($p = .856$)

[11] Again, note that this value will not match published values exactly. That is an artifact of the way that we have simulated data for the online course resources to allow students to practice the analysis, not a commentary on the published results.

[12] This, again, will not match the published results exactly due to the simulation process for the practice data in the online resources.

or the seven-week follow-up (p = .970). Finally, there was no significant difference between the three-week and seven-week follow-up (p = .986). Compared to pre-test knowledge scores (M = 2.490, SD = 1.510), scores were higher at post-test (M = 5.580, SD = 1.610), the three-week follow-up test (M = 5.840, SD = 1.610), and the seven-week follow-up test (M = 5.720, SD = 1.730). These results suggest increases in knowledge about racial microaggressions following the workshop that remained relatively stable at three- and seven-week follow-ups.

FUTURE CASES

While we hope that this resource is helpful, we know that it is incomplete and represents a small cross-section of the available research on race, racism, anti-Blackness, white supremacy, and whiteness in education. We encourage you to use your library resources, textbooks, databases, and online journals to search for, critique, and understand other examples of research in this area.

We are grateful to Routledge for allowing us to make this an online, free resource. Another advantage of this format is that we can expand these cases over time. As you continue exploring the existing research and conducting your own, we would be thrilled to hear about more published studies that could be examples for a resource such as this. We also hope to expand the types of cases included in these resources to more broadly defined issues of equity and justice in education. If you identify other published studies, we should consider including in this and other resources, let us know! You can submit the details at https://aub.ie/casestudies

REFERENCES

Banks, B. M., Adams, D. F., Williams, C., & Piña, D. (2020). Preliminary investigation of efforts to improve awareness of racial microaggressions on campus. *Journal of Underrepresented and Minority Progress, 4*(1), 20-43. https://doi.org/10.32674/jump.v4i1.1763

Cisneros, J., & Cadenas, G. (2017). DREAMer-ally competency and self-efficacy: Developing higher education staff and measuring lasting outcomes. *Journal of Student Affairs Research and Practice, 54*(2), 189-203. https://doi.org/10.1080/19496591.2017.1289098

Fitchett, P. G., Starker, T. V., & Salyers, B. (2012). Examining culturally responsive teaching self-efficacy in a preservice social studies education course. *Urban Education, 47*(3), 585-611. https://doi.org/10.1177/0042085912436568

Garvey, J. C., Mobley, S. D., Summerville, K. S., & Moore, G. T. (2019). Queer and trans* students of Color: Navigating identity disclosure and college contexts. *Journal of Higher Education, 90*(1), 150-178. https://doi.org/10.1080/00221546.2018.1449081

Harber, K. D., Reeves, S., Goman, J. L., Williams, C. H., Malin, J., & Pennebaker, J. W. (2018). The conflicted language of interracial feedback. *Journal of Educational Psychology, 111*(7), 1220-1242. https://doi.org/10.1037/edu0000326

Kumar, R., & Hamer, L. (2013). Preservice teachers' attitudes and beliefs toward Student Diversity and Proposed Instructional Practices: A Sequential Design Study. *Journal of Teacher Education*, *64*(2), 162–177. https://doi.org/10.1177/0022487112466899

Marcucci, O. (2020). Implicit bias in the era of social desirability: Understanding antiblackness in rehabilitative and punitive school discipline. *Urban Review: Issues and Ideas in Public Education*, *52*(1), 47–74. https://doi.org/10.1007/s11256-019-00512-7

Matthews, J. S., Kizzie, K. T., Rowley, S. J., & Cortina, K. (2010). African Americans and boys: Understanding the literacy gap, tracing academic trajectories, and evaluating the role of learning-related skills. *Journal of Educational Psychology*, *102*(3), 757. https://doi.org/10.1037/a0019616

McGowan, S. L., & Kern, A. L. (2014). Pre-service foreign language teachers' attitudes of privilege and oppression. *Journal of Education and Training Studies, 2*(1), 31-43. https://doi.org/10.11114/jets.v2i1.188

Saleh, M. F., Anngela-Cole, L., & Boateng, A. (2011). Effectiveness of diversity infusion modules on students' attitudes, behavior, and knowledge. *Journal of Ethnic &*

Cultural Diversity in Social Work, *20*(3), 240–257. https://doi.org/10.1080/15313204.2011.594995

Strunk, K. K., & Locke, L. A. (Eds.) (2020). *Research methods for social justice and equity in education.* Palgrave.

Strunk, K. K., & Mwavita, M. (2020). *Design and analysis in educational research: ANOVA designs in SPSS.* Routledge.

Wasserberg, M. J. (2014). Stereotype threat on African American children in an urban elementary school. *Journal of Experimental Education, 82*(4), 502-517. https://doi.org/10.1080/00220973.2013.876224