“Our strength can come through multiplicity rather than adherence to a single norm.” (Riley, 2017, p. 263)

A doctoral student is reading an engineering education research paper addressing a topic related to her work. She finds a graph that reports differences in students’ grades in multiple sections of the same class taught by different instructors. The graph separates sections based on whether they received an intervention (treatment) or did not (control), and they report a p-value less than 0.05 based on comparing final exam grades to support the effectiveness of the intervention. The graph does not explain how many instructors are in each group, how the instructors were selected or the length of the intervention, so the doctoral student questions the legitimacy of the research claims.

In this editorial, we are taking up Riley’s challenge to the field to move beyond “a single norm” (2017) and explore ways quantitative analysis can be used in richer and more nuanced ways in educational research. The American Statistical Association (ASA) and the American Educational Research Association (AERA) recently sponsored the ASA Symposium on Statistical Inference entitled, “Scientific Methods for the 21st Century: A World Beyond p < 0.05” (August 2017). The argument was made during the conference that the p-value is overused in educational research and is problematic for the field in terms of how it is interpreted and used to describe results and implications of studies. Professional journals (e.g., Basic and Applied Psychology) and statisticians have discussed the issues of using p-values as the sole measure for significant results and the implications for research (e.g., Nuzzo, 2014; Trafimow & Marks, 2015), and the ASA has also released a statement (discussed later in this paper) that includes principles for considering the p-value as a measure for statistical inference (Wasserstein & Lazar, 2016). As illustrated in the vignette and as argued here, a significant p-value alone is not enough to support a research claim. This editorial addresses how we as researchers can move beyond the p-value to more deeply understand educational research and the arguments that can be made using quantitative data as evidence.

The fundamental goal of quantitative research is to make a convincing argument based on numerical data in response to a research question. An important component is recognizing that this is a process of modeling phenomena using evidence that has been mapped onto a quantitative framework for analysis (Sloane & Wilkins, 2017) as the data are not numeric in their natural state. Instead, a researcher imposes a numerical map on them (Becker, 2017). This process of modeling a phenomenon is a decision-making process wherein decisions need to be documented (Ercikan & Roth, 2006; Sloane & Wilkins,
Since different researchers are likely to make different decisions, the documentation process about how and why the numerical map was created is essential. Users or interpreters of the research results need to consider the effects that a slightly different approach or decision might have on the results of the investigation.

When deciding to model data numerically, the following questions should be considered:

1. What is the research question? What information is necessary to appropriately address this question?

2. What is the nature of the data to be collected (e.g., how, why, when, and from whom)? Can these data be modeled numerically?

3. What is the theory concerning the underlying phenomenon that is being used to design the study? How well does the model align with the theory?

4. How are the data collected? How are groups assigned? How can quantitative values be assigned to the data (e.g., what measure is being used)?

5. What are limitations to the study and how it was conducted?

**Aligning Research Questions to Data Collection and Analysis**

Fundamentally, data collection and analysis are guided by the investigator’s research question and the nature of the phenomenon to be explored (Ercikan & Roth, 2006). Some questions lend themselves to quantitative or statistical analyses while others to qualitative or detailed examinations of data. Still other questions can best be addressed through a combination of the two types of analysis known as mixed methods. It is the research question or set of questions that directs the researcher to consider whether the research design calls for quantitative analysis (U.S. Department of Education & National Science Foundation, 2013). We explore here the kinds of research questions that lend themselves to quantitative analysis. Later editorials will address considerations for qualitative research and mixed methods research.

There are two primary types of quantitative questions: comparative or relationship. Comparative research questions in engineering education might be, “Are the performance scores of electrical engineers the same as mechanical engineers on the Fundamentals of Engineering Exam upon graduation?” or “Do the entering SAT scores of computer engineers exceed 1300 at a given institution?” In both cases, numerical comparisons are being made between two quantities.

Relationship research questions could include “How are the SAT scores of an entering biological engineering student related to their exiting GRE score?” or “How do female attitudes toward STEM relate to engineering degree completion?” In both of these questions, a relationship is being examined between two factors. It is important to note that the results of a relationship study do not imply cause and effect. There are additional design features that are necessary to control for extraneous variables when examining causal relationships.

Once a quantitative research question is formed, many researchers select a data collection process (e.g., using a survey or assessment), select and implement a statistical analysis, analyze the p-value, and draw a conclusion. The t-test is often selected for statistical analysis because it is the first test researchers learned as students and, hence, the one they are most comfortable with. However, this is not an appropriate approach for selecting an analysis technique. The researcher should examine the nature of the phenomenon, the question, the data, and then determine an appropriate statistical model.
Assumptions of a Statistical Test: t-test

Many educational researchers learn about the Student t-test as it is the most common introductory statistical analysis. Students typically learn the difference between forms of the t-test and practice applying these techniques to problem sets designed to match its assumptions. Because students do not have to evaluate the appropriateness of the t-test or their calculations while learning this technique, they may develop the inappropriate belief that statistical analysis entails calculating and applying a p-value, regardless of the underlying assumptions of the statistical tests and the nature of the underlying phenomenon. In many cases, students learn additional statistical techniques in the same sterile environment, often without the knowledge of underlying assumptions that should be verified before applying each test. Statistical tests, whether parametric (viz., based on an assumed population distribution) or non-parametric (viz., not based on an assumed population distribution), always have underlying assumptions. The researcher is responsible for collecting data in a manner that conforms to the assumptions of the test and for examining whether the data do indeed conform to the assumptions of the test.

While the t-test has multiple versions, we limit this discussion to the t-test for independent samples. Once independence of the two samples is confirmed, there are three additional assumptions: the data follow either a continuous or an ordinal scale, the data resulted from a random sample of the total population and the data are drawn from a normal distribution or from a sufficiently large sample that its distribution of means follows a normal distribution. For selecting an appropriate sample size, researchers should consider using power analyses to determine the appropriate sample size for the goals of the study. Power refers to the probability of correctly rejecting a hypothesis that is false.

The first assumption refers to the nature of the data. Continuous data can be measured at a desired level of accuracy. Ordinal data are categorical but also have a defined order. A common example of ordinal data is a Likert scale. Strongly disagree, disagree, agree, and strongly agree can be mapped to the numerical values of 1 to 4, with order indicating the level of agreement. In short, the researcher should consider how the data are being mapped to a quantitative model as evidence to support results and conclusions.

The next assumption is that the data were drawn from a random sample of the total population. This is rarely the case in educational investigations. For example, we will assume the researchers want to compare the average performance for students who do and do not receive a certain treatment. If all assumptions are met, including random placement, an unpaired t-test could be used. But random assignment is unlikely in this situation; students are humans and humans have the option of selecting in or out of a treatment. With self-selection and choice, the assumption of randomness is violated.

The third assumption of the t-test is that the data are drawn from a normal distribution or the sample size is large enough that the distribution of means approaches normality. For small sample sizes (typically defined as less than 30), the original data need to be examined to determine whether they are likely to have been drawn from a normal distribution. Thankfully, most statistical packages have built in nonparametric data analysis tools to assist in verifying this assumption. A common test for normality is the Shapiro-Wilk (Ghasemi & Zahediasl, 2012). Monte Carlo simulations can also be used to help define appropriate sample sizes (Muthén & Muthén, 2002).

As this example illustrates, there are many ways in which the data may not conform to the assumptions of a t-test. Even if a research question requires a comparison between two
means, a t-test may not be appropriate because of these violations. Most statistical tests, parametric or non-parametric, have a set of assumptions that should be analyzed to determine the fit of the model to the data before applying the analysis and interpreting the p-value. If the assumptions of a model do not fit the data, then interpreting the resultant p-value is inappropriate.

**Resultant p-values and Their Interpretation**

The intention of this discussion is not to eliminate the use of p-values in quantitative educational research. Rather, it is to develop a better understanding of how and when to use p-values and what other information should be provided to substantiate claims and provide a rich description of the study. The ASA has identified the following six principles for the use of p-values (Wasserstein & Lazar, 2016, pp. 131–132):

1. p-values can indicate how incompatible the data are with a specified statistical model.
2. p-values do not measure the probability that the hypothesis studied is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value or statistical significance does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

These principles imply that additional information is needed beyond a p-value, with some authors recommending that effect sizes and confidence intervals should also be reported (Cumming, 2008; Nuzzo, 2014). Effect size quantifies the differences between the two groups, allowing the researcher to determine whether the difference has practical significance, for example weighing a small effect size against the expense of a given intervention. A confidence interval provides a range of values for a given mean or difference between means, and the researcher can be confident (at a desired percentage of confidence) that the actual value lies within that range. A confidence interval more clearly conveys the possibility of a Type I error (incorrectly rejecting a true hypothesis) than does a p-value. Other authors suggest including information about the distribution of the data (e.g., via standard deviation or graphical representations like boxplots) to better represent the phenomenon of interest and identify patterns for further investigation (Valentine, Aloe, & Lau, 2015).

In addition, the variability in the data can be more interesting in educational phenomenon than the average performance or even the difference in average performances. For example, in a nationwide study of Calculus instruction that used a quantitative analysis to identify schools with exceptionally high calculus performance (Bressoud, Carlson, Mesa, & Rasmussen, 2013; Bressoud & Rasmussen, 2015), understanding the outliers or high performers was as important as understanding the averages. Another example is the use of pre- and post-intervention assessment scores to examine learning. When pre- and post-assessments and average difference scores are used in isolation, learning is treated as a linear
process between two points in time. Student learning, however, may move in a circular or step-function manner over time, or continue beyond the final measurement. Reporting a p-value alone based on pre-post assessments fails to capture the context of the study or the theoretical model that the researchers might use to interpret the evidence. In general, researchers should seek descriptions that capture a richer sense of the phenomenon being investigated. This may be reflected through the models that they select, the analyses they complete and the detailed descriptions they provide.

**Experimental and Quasi-Experimental Designs**

As Becker describes reporting research, “Social scientists combine three components – data, evidence, and ideas (sometimes called ‘theories’ or ‘concepts’) – to convince themselves, their colleagues, maybe even a wider audience, that they have found something true, something more than a coincidence or an accident” (2017, p. 4). The overarching goal when reporting research is creating an argument to support the claims. In experimental designs, the researcher would like to investigate whether the treatment (and not some other variable) has had an impact on the outcomes. In true experimental designs, the treatment and control groups are randomly determined, meaning all participants have the same probability of being placed in either group. Randomness helps protect unintended variables from influence on the implementation or the outcomes. However, educational settings often make randomness impossible to achieve.

Many educational researchers use quasi-experimental designs when random assignment is not possible. A challenge in implementing such a design is clearly defining the treatment and the control. Quasi-experimental designs use many of the same parametric and nonparametric statistics for analyzing data as do true experimental designs. However, data must be interpreted with careful consideration that the assumption of random selection has been violated. In quasi-experimental designs, the researcher is aware of the violations of the assumptions and completes additional analyses to establish the comparability of the groups and examine the internal validity of the conclusions. Internal validity refers to whether there are alternative explanations or factors that would explain the outcomes of the experiment beyond those under investigation. Researchers should also report the groups’ characteristics to explain how they are comparable based on the variables of interest. For example, when comparing two semesters of a course, attributes such as students’ majors or math backgrounds should be similar between the two semesters.

Systematically investigating alternative explanations can lead to additional quantitative or qualitative investigations into the phenomena under investigation as well as investigations into the variables that threaten the validity of the conclusions. For instance, if the treatment is teaching a laboratory in a particular manner, observation (a qualitative method) can be used to determine whether the instructors adhered to the prescribed instructional approach. Observation can also be used to determine whether the treatment and control conditions were implemented as planned. Triangulation, or the use of multiple investigative methods into the same phenomenon, builds evidence throughout the investigation and typically strengthens conclusions.

**Concluding Remarks**

There is nothing intrinsically wrong with p-values, but the p-value does not stand alone. The challenge for educational researchers is providing enough information surrounding the
p-value to create a compelling argument to support a research claim. In addition, researchers should consider different types of quantitative analyses and measures to present a richer description of the phenomenon. Educational researchers need to be aware of both the limitations of p-values and the complexities of the educational research process due to the multifaceted nature of studying human subjects. They also need to have a firm grounding in educational research techniques and their limitations. This may require creating research teams that include experts in various methodologies such as quantitative analysis of educational data for research and evaluation. In addition, educational researchers should also attend to the changing landscape of quantitative data that may be available over time and with advanced technologies.

References

Quality Considerations in Education Research


**Authors**

Margret Hjalmarson is Associate Professor in the College of Education and Human Development at George Mason University, Fairfax, VA 22030; mhjalmar@gmu.edu.

Barbara Moskal is a Research Professor in the College of Education at Texas Tech University, Lubbock, TX 79409; barbmoskal@gmail.com.