Learning Disabilities Research & Practice, *32*(4), 208–215

DOI: 10.1111/ldrp.12147

Null Hypothesis Significance Testing and *p* Values

Jason C. Travers
University of Kansas
Bryan G. Cook and Lysandra Cook
University of Hawaii

p values are commonly reported in quantitative research, but are often misunderstood and misinterpreted by research consumers. Our aim in this article is to provide special educators with guidance for appropriately interpreting *p* values, with the broader goal of improving research consumers' understanding and interpretation of research findings. Specifically, we discuss null hypothesis significance testing, describe what *p* values mean and how they are reported, describe some common misconceptions of *p* values, and provide two examples from the research literature to illustrate how *p* values are used in the field. Our take-home message is that *p* values indicate how likely study results are to occur if the null hypothesis is true, and that *p* values should be cautiously interpreted.

Ms. Freiheit is a literacy specialist who receives two journals each month as a membership benefit in a professional association. One journal primarily publishes research, which she finds interesting but sometimes has difficulty understanding. She read a study of an oral reading fluency intervention that repeatedly referred to results as statistically significant when $p < .05$. To better understand the meaning of the results, Ms. Freiheit turned to a group of special educators on an online forum. She asked a question about the meaning of statistical significance and p values. Answers were abundant but contradictory, which confused Ms. Freiheit and resulted in a debate on the online forum. Some remarked that smaller p values meant that the study was higher quality and proved that the intervention worked. Others suggested that she pay attention to effect sizes, and that p values didn't matter very much. A back-and-forth conversation between two teachers became adversarial and personal. Exasperated and unsure what to believe, Ms. Freiheit turned off notifications to her post and later deleted the question.

An evidence-based approach to special education depends largely on the research knowledge of professionals who teach and serve students with disabilities. Savvy professionals turn to the research literature for practical guidance, but critically evaluating a study and interpreting its findings is a daunting task. One metric commonly relied on when interpreting research findings is the *p* statistic or *p* value. When consumers of research see a *p* value reported (e.g., $p < .05$), they may know it indicates something important about the significance of study findings, but often aren't sure exactly what it means. Educators often infer that a low *p* value indicates the study was successful (e.g., the intervention examined was effective), and/or the findings are important. But these conclusions are not technically accurate. Indeed, despite ubiquitous reporting of *p* values in quantitative analysis, both researchers and research consumers commonly misunderstand the *p* value (Goodman, 2008; Hubbard & Lyndsay, 2008; Wasserstein & Lazar, 2016). Misunderstandings about the *p* value are commonplace and can mislead professionals to wrongly believe that statistically significant findings prove that an intervention was effective.

The *p* value, hypothesis testing, and statistical modeling are technical concepts and, in some cases, can be exceptionally confusing. In this article, we seek to clarify what *p* values do and do not mean, and provide guidelines for appropriately interpreting *p* values in educational research. Our aim is to provide guidance to special educators seeking to evaluate research results. We provide some preliminary and general discussions of *p* values, which are admittedly incomplete, with the goal of clarifying this abstract and sometimes contentious statistic to special educators who are not experts in statistical analysis.

Throughout this article, we discuss *p* values and related concepts by referencing group research. Although some single-case design researchers have begun to incorporate statistical analyses into the methodology and occasionally use *p* values to inform some conclusions about their data, consensus about this approach to data analysis has yet to be reached. Thus, although research consumers may observe statistical data and *p* values in published reports of single-case experiments, we do not address this topic because it is currently an unresolved and emerging area. Instead, our focus is on the *p* value as it relates to traditional group research, especially group experimental research—something that special educators often consider when examining the effectiveness of instructional practices (Cook & Cook, 2016).

Our primary message is that the *p* value indicates how likely study results are to occur if the null hypothesis (e.g.,

Requests for reprints should be sent to Jason C. Travers, University of Kansas. Electronic inquiries should be sent to jason.travers@ku.edu.

LEARNING DISABILITIES RESEARCH & PRACTICE

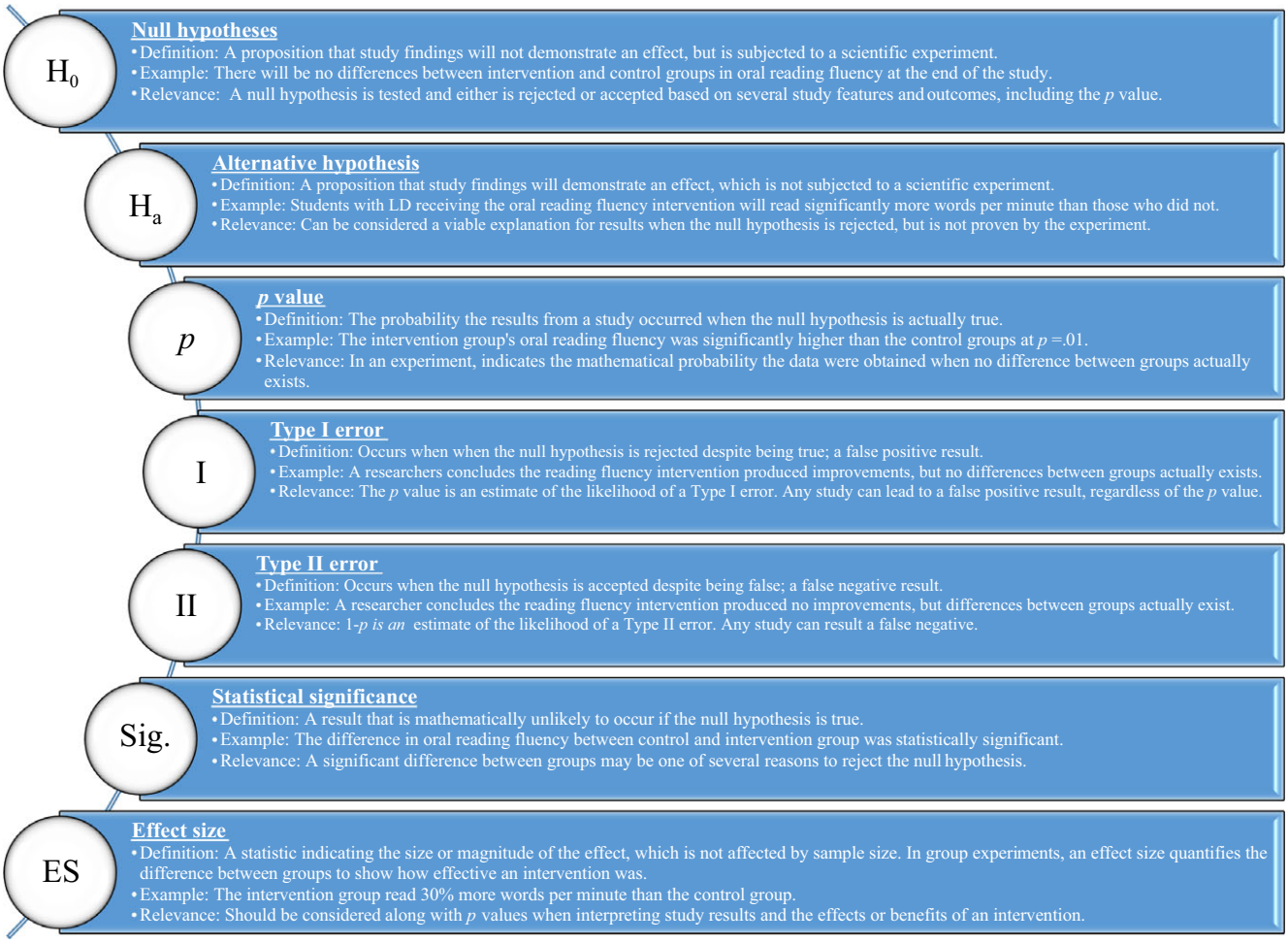

FIGURE 1 Definitions, examples, and relevance of key concepts. [Color figure can be viewed at wileyonlinelibrary.com]

an intervention will not be effective for a population of learners) is true, and that *p* values should be cautiously interpreted. To make sense of what *p* values mean, one must understand several terms and concepts associated with group research (see Figure 1) as well as the role of hypotheses, and null hypotheses in particular, in scientific research. After a discussion of null hypothesis significance testing (NHST), we provide an overview of what *p* values mean, as well as some prominent misconceptions about this common statistic. We also review the use of *p* values in two recent intervention studies in special education to illustrate how special educators can appropriately interpret *p* values when they read experimental research to further their professional development and practice.

NULL HYPOTHESIS SIGNIFICANCE TESTING

Hypotheses are propositions or conjectures made on the basis of incomplete information; in other words, an educated guess. By hypothesizing about what they will find before they conduct the study, researchers provide testable predictions that are a critical part of the scientific method. In research reports, authors often state one or more hypotheses at the end of the introduction section. As an example, assume a researcher believes, based on theory and previous research that an intervention is likely to result in improved oral reading fluency for students with learning disabilities (LD). Therefore, before conducting an experiment to examine the efficacy of the intervention, the researcher hypothesizes that the intervention will cause improved oral reading fluency for students with LD.

It is often assumed that science establishes knowledge by researchers conducting a study (or multiple researchers conducting multiple studies) that proves a research hypothesis (e.g., that a particular intervention improves oral reading fluency for students with LD). However, the way hypotheses function in modern scientific research is a bit more complicated. Fisher (1925) and Popper (1959) argued that one cannot definitively prove a theory by positively affirming it through research, but one can clearly falsify a theory by refuting it. Using Popper's classic example to illustrate this point, no matter how many times a hypothesis that all swans are white is confirmed through empirical research (e.g., observing swans), one cannot definitively prove that all swans are white because researchers cannot observe all possible swans.

In other words, a black swan that researchers have not yet discovered may exist. In contrast, just one instance of falsification (e.g., sighting a black swan) refutes the proposition. Similarly, no matter how many times an oral reading fluency intervention is found to be effective for students with LD, we cannot conclude that all students with LD will benefit from the intervention. This is because researchers are unable to examine the effects of the intervention for all students with LD. Because one can refute a theory by falsifying a hypothesis, but can never definitively prove a theory no matter how much empirical support is gathered, most quantitative scientific research is designed to test, and potentially falsify, a hypothesis that is contrary to the researcher's theory. Thus, the null hypothesis is the basis of NHST and the referent of *p* values.

In our example of a researcher who theorizes that an intervention will improve oral reading fluency for students with LD, the null hypothesis (i.e., the hypothesis that will be nullified if study findings support the researcher's theory) could be stated as *the intervention results in no difference in the oral reading fluency for students with LD*. In other words, the null hypothesis, which is often represented as H_0, states what will occur if the researcher's theory is not supported by the research. In intervention research, the null hypothesis typically predicts that an instructional practice will have no effect, or that there will be no differences in performance between the learners who do and do not receive the intervention. The alternative hypothesis rivals the null hypothesis and predicts what results will occur if the researcher's theory is valid. In our example, the alternative hypothesis could be stated as *the intervention improves oral reading fluency for students with LD*. Though the alternative hypothesis, typically represented as H_1 or H_a, corresponds with the researcher's theory-based prediction, it is not actually being tested during an experiment. Rather, researchers test the degree to which study results are consistent with the null hypothesis. If the null hypothesis is rejected, then the alternative hypothesis may be indirectly supported. Each time a similar study refutes the null hypothesis, support for the alternative hypothesis grows incrementally stronger, even though the alternative hypothesis is never completely proven.

NHST is a conservative approach to accumulating scientific knowledge. The assumption is that the null hypothesis is true until it is sufficiently refuted. This approach stands in contrast to pseudoscience, in which a theory or practice is assumed valid until proven otherwise (Travers, 2017). This approach is analogous to the U.S. criminal justice system, in which defendants are assumed innocent until proven guilty beyond a reasonable doubt. In a sense, the null hypothesis in criminal cases is that the defendant is not guilty. Just as criminal courts are designed to err on the side of not falsely convicting innocent people, NHST errs on the side of not identifying practices as effective until convincing evidence has refuted the null hypothesis.

Identifying the null and alternative hypotheses in research articles often is challenging. Although researchers sometimes state the null and alternative hypotheses, they more typically just state the alternative hypothesis, or only list one or more research questions without specifying hypotheses. Although it would be helpful for researchers to formally state their null and alternative hypotheses, hypotheses can typically be inferred without much difficulty. For example, the author of our example study might have just stated the following research question: Does the intervention result in improved oral reading fluency for students with LD? Although unstated, one can infer the null hypothesis is that the intervention will not improve oral reading fluency, and the alternative hypothesis is that it will.

After research questions are posed and hypotheses have been formulated, the researcher designs a study to test the null hypothesis. In group experimental studies, groups of participants are selected and assigned to control and intervention conditions. Assessments typically are conducted to determine prestudy performance, and to ensure the groups are not different in ways that might affect the results. The intervention and control phases are instituted and, after the intervention is completed, researchers again administer assessments to both groups of learners. At this point, the experiment is completed, but many steps of the study remain. The researchers must now analyze their data to determine whether sufficient evidence allows rejection of the null hypothesis. One piece of evidence that helps to determine whether to reject the null hypothesis is the *p* statistic.

WHAT THE *p* VALUE MEANS

Because no study is perfect and all studies involve error, researchers may design experiments that detect no true effects, but wrongly conclude the intervention was effective (i.e., a false positive finding; Type I error) on the basis of study findings. Conversely, researchers may develop an effective intervention, but then conduct a study that fails to detect the effects and wrongly conclude the intervention was not effective (i.e., a false negative finding; Type II error). The *p* value can help researchers draw reasonable conclusions about a study's results by indicating how likely their findings occurred when the intervention truly had no effect. In other words, the *p* value (e.g., $p = .03$) indicates the probability of making a Type I error by rejecting the null hypothesis.

p values are reported as decimals that often, but not necessarily, are rounded to two decimal places and are interpreted as probability percentages. For example, $p = .03$ means that there is a 3 percent probability that the study results occurred as they did if the null hypothesis is true. Larger *p* values (e.g., $p = .60$) indicate a strong probability that results occurred as they did with a true null hypothesis (i.e., findings are relatively consistent with the null hypothesis), and lower *p* values (e.g., $p = .01$) indicate a small probability that results occurred as they did with a true null hypothesis (i.e., findings are relatively inconsistent with the null hypothesis). *p* values range between 0 and 1. Because there is never a 0 percent probability that results occurred as they did with the null hypothesis being true, a *p* value is never zero. Similarly, because results never absolutely prove the null (or any other) hypothesis, a *p* value is never 1.

Recall our scenario of an experimental study being conducted to examine the effect of an intervention on oral reading fluency. Assume the oral reading fluency of participants

in the experimental group (who received the intervention) improved quite a bit more over the course of the study than for participants in the control group (who did not receive the intervention). Even though these findings appear inconsistent with the null hypothesis, there is still some possibility that differences between the groups occurred in the study even though the null hypothesis is true. For example, it is possible that more "ready-to-learn" emergent readers, who were just starting to make large gains in reading performance when the study started, were randomly assigned to the experimental group. In this scenario, although the intervention group demonstrated a large improvement in oral reading fluency, their gain had little or nothing to do with the intervention. Therefore, studies showing the experimental group improved more than the control group may reflect sampling problems and not the effects of the intervention. In this case, the null hypothesis (i.e., the intervention results in no difference in oral reading fluency for students with LD) is true for the broader population, even though study findings showed the experimental group outperformed the control group (i.e., a Type I, false positive error). Because some error is present in every study, and especially in studies conducted in applied settings like schools and classrooms, there is always some possibility the null hypothesis is true, no matter what the study findings are.

Three factors impact the *p* value: effect size (i.e., magnitude of effect), variability, and sample size. Using our example of a reading intervention study, the *p* value is influenced by (a) the magnitude of the difference between groups on the outcome measure (oral reading fluency), (b) the variability in performance on the outcome measure within each group, and (c) the number of participants in the study. All else being equal, stronger evidence against the null hypothesis (and a lower *p* value) is associated with the experimental group outperforming the control group by a lot rather than a little. Lower variability within groups also is associated with lower *p* values. For example, a difference between groups of, say, seven words per minute is inconsistent with the null hypothesis if all students in the experimental group improved between 12 and 14 words per minute ($M = 13$) whereas all participants in the experimental group improved between 5 and 7 words per minute ($M = 6$). However, those same mean values are less inconsistent with the null hypothesis if student performance within groups is highly variable, with some students in the experimental group making no progress at all and some students in the control group making huge gains. Finally, all else being equal, studies with large numbers of participants have lower *p* values. The likelihood that differences between groups are due to sampling error (and not due to the null hypothesis being false) decreases with large samples. For example, the performance of a few participants with idiosyncratic learning responses (i.e., outliers) in a treatment condition will have little impact on the mean performance of a group in a study with 200 participants per group. Thus, for group intervention studies, small *p* values are most likely to occur in studies with large differences between groups, small variability within groups, and large samples.

How small should a *p* value be to conclude that the null hypothesis should be rejected? Fisher (1925) noted that $p = .05$ (i.e., a 5 percent likelihood that results occurred as they did if the null hypothesis is true) is a convenient criterion for denoting interesting, or significant, findings that warrant rejection of the null hypothesis—a figure that remains the common standard. Although $p = .05$ means the null hypothesis will be falsely rejected (i.e., Type I error) in one of every twenty analyses (5 percent), using a lower *p* value as the criterion for statistical significance would increase the likelihood of accepting the null hypothesis when it should be rejected (i.e., Type II error). Many researchers and research consumers have become accustomed to $p < .05$ signifying that results are statistically significant, but there is nothing inherently important about that *p* value; it is an arbitrary criterion that has become ingrained over many years. Reflecting the commonly accepted approach of using a *p* level of .05 as the threshold for statistical significance, some researchers simply report whether *p* values are $<.05$ (and statistically significant) or $>.05$ (and not statistically significant), rather than reporting the specific *p* value (e.g., $p = .034$). One might also see researchers reporting $p < .01$ or $p < .001$ to indicate very low *p* values. We agree with Rosnow and Rosenthal's (1989) sentiment about *p* values that "surely, God loves the .06 nearly as much as the .05" (p. 1277). Indeed, *p* values should be thought of as a continuum of probabilities, not as a strict dichotomy of statistical significance. Additional internet resources that special education professionals may find helpful for understanding *p* values and NHST can be found in Figure 2.

WHAT THE *p* VALUE CANNOT TELL YOU

Unfortunately, inaccurate beliefs about what the *p* value means have proliferated for generations, and many researchers and research consumers subscribe to one or more erroneous beliefs about this ubiquitous statistic. In this section, we present several common misconceptions related to *p* values. This is not intended to be an exhaustive discussion of the issues and problems related to *p* values. We refer interested readers to several detailed overviews for additional information (e.g., Goodman, 1993, 2008; Greenland et al., 2016; Hubbard & Lyndsay, 2008; Sullivan & Fein, 2012).

A Low *p* Value Indicates Practical Significance

This is arguably the most prominent misconception, and illustrates well the difficulty associated with interpreting *p* values. Importantly, when applied to group experiments the *p* value can inform research consumers about the degree to which differences between groups are statistically interesting, but it cannot inform conclusions about the practical significance of a difference. Practical significance, sometimes referred to as the "clinical significance" or "social validity" of results, refers to whether a difference between groups has any practical or real-world value. This is important because sample size directly affects *p* values, and large samples can cause relatively small (i.e., practically useless) differences between groups to be statistically significant. Thus, a low *p* value

Web pages and articles

- Five guidelines for using *p* values: http://blog.minitab.com/blog/adventures-in-statistics-2/five-guidelines-for-using-p-values
- How to correctly interpret *p* values: http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-correctly-interpret-p-values
- Not even scientists can easily explain *p* values: http://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/
- What a *p* value tells you about statistical data: http://www.dummies.com/education/math/statistics/what-a-p-value-tells-you-about-statistical-data/

Videos

- Hypothesis testing and *p* values by Khan Academy: https://youtu.be/-FtlH4svqx4
- Is most published research wrong? https://www.youtube.com/watch?v=42QuXLucH3Q&t=26s
- Understanding *p* values – Statistics help: https://www.youtube.com/watch?v=eyknGvncKLw
- What is a null hypothesis (and alternate hypothesis): https://www.youtube.com/watch?v=tDmCFVQvv2A
- What makes science true?: https://www.youtube.com/watch?v=NGFO0kdbZmk

Podcasts

- One humble test that makes or breaks companies and careers: https://www.statnews.com/2017/03/31/pvalue-statistics-podcast-science
- *p* values: https://itunes.apple.com/us/podcast/mini-p-values/id890348705?i=1000314963010&mt=2

FIGURE 2 Videos, podcasts, and webpages for learning more about *p* values and null hypothesis statistical testing. [Color figure can be viewed at wileyonlinelibrary.com]

cannot inform conclusions about the real-world importance of the difference between groups. For example, our oral reading fluency study might have 500 participants in the intervention and control conditions—1,000 total participants. Statistical analysis of study results might reveal a difference between groups that is statistically significant (e.g., $p = .001$). However, on average the intervention group only read three words per minute (103 words per minute) more than the control group (100 words per minute)—an effect size of about 3 percent. Although the difference is statistically significant, most educators would consider this outcome to have little practical significance because the small improvement does not justify investment in the intervention.

Professionals should be cautious about drawing practical conclusions based solely on a low *p* value, especially if the study included a large sample size. Although a large sample size is important for drawing inferences about the generalizability of the findings (Cook & Cook, 2017), the *p* value cannot inform conclusions about the practical value of the findings. Professionals should rely on effect sizes to determine the practical importance of study findings. The effect size informs decisions about "the magnitude of the difference between groups . . . (and) is the main finding of a quantitative study" (Sullivan & Fein, 2012, p. 279). The effect size can be reported in several different ways depending on the statistical method of analysis, but often is reported along with interpretations as to whether the effect is small, moderate, or large. In most cases, professionals will be interested in the size of the effect (i.e., the amount of change associated with an intervention) rather than the mere presence of a statistically significant effect (i.e., the *p* value).

A Low *p* Value Means it is Correct to Reject the Null Hypothesis

Goodman (1993) explained why the *p* statistic alone is insufficient reason to reject the null hypothesis. Researchers and research consumers should examine other features of the study—including participant selection, assignment of participants to treatment and control conditions, adherence to intervention procedures, and attrition of participants from the study—to justify rejection of the null hypothesis. Careful examination of the data and statistical methods used also is necessary to ensure analyses were conducted in ways that do not produce erroneous results (Nuzzo, 2015). Finally, other explanations that might reasonably account for observed differences between the groups should be examined. If the study was conducted appropriately and no competing explanations are found, then the collective evidence from the study, including a low *p* value, may justify rejection of the null hypothesis. In essence, results from poorly designed or conducted studies should not be used to justify rejecting the null hypothesis regardless of the *p* value.

A Low *p* Value Indicates the Likelihood the Alternative Hypothesis is True

Evidence that the null hypothesis is incorrect does not constitute direct evidence that the alternative hypothesis is true. Many individuals, including researchers, may be under the impression that obtaining a low *p* value and rejecting the null hypothesis means "the intervention worked." This assumption is not accurate. Evidence that the null

hypothesis is incorrect does not qualify the alternative hypothesis as true because only the null hypothesis is actually tested via experimentation. Although deductive logic may lead to conclusions that an intervention is the most plausible explanation for differences between groups, this claim (i.e., the alternative hypothesis) is not put to a direct, scientific test. Analogously, concluding insufficient evidence exists to convict someone of a crime is not the same as concluding the person is innocent. The former claim is the product of insufficient evidence (of guilt), and the latter requires presenting evidence to establish innocence (a different claim). Similarly, rejecting the null hypothesis from our oral reading fluency experiment does not mean that our intervention caused the treatment group to perform better than the control group. Many alternative explanations could be the cause for the observed differences between groups (e.g., sampling error). We must look to the *p* value *and other factors* (e.g., whether the intervention procedures were applied as designed, or with fidelity) to decide whether the alternative hypothesis can be considered a viable explanation for our results. Moreover, results from a single study, or from a few studies, should not be considered reason to believe an intervention will have the intended effect for the broader population regardless of the *p* value. Instead, initial rejection of the null hypothesis signals the need for additional research. If a sufficient number of high-quality replication studies generally find the same results, then confidence about the positive effects of the intervention is warranted (Travers, Cook, Therrien, & Coyne, 2016).

A Low *p* Value Indicates the Likelihood Results Occurred by Chance

As explained previously, the *p* value represents the likelihood results would have occurred if the null hypothesis was true. A common but mistaken belief is that the *p* value informs researchers how likely it is that their results were caused by chance alone. In other words, a *p* value of .05 may be wrongly perceived as evidence that results from an experiment have only a 5 percent probability of being the product of coincidence (and are 95 percent likely to be caused by the intervention). But the *p* value does not show the researcher why the results were statistically significant and, consequently, the *p* value cannot be used to conclude whether the results occurred due to coincidence or any other reason. Thus, the *p* value can tell us the degree to which results are inconsistent with the null hypothesis and worthy of further consideration, but cannot tell us why the results were obtained, or how likely it is that results occurred by coincidence. Inaccurate interpretations like these can lead to overconfidence about the results of a single study or body of research.

A Low *p* Value Indicates the Likelihood Results are Generalizable

Individuals may presume that a low *p* value indicates the results are generalizable, or are more likely to apply to other individuals in the population (e.g., all students with LD), and that studies with smaller *p* values are more generalizable than studies with larger *p* values. These presumptions are not accurate. Generalizability of results is dependent on several factors, with sample representativeness being the primary concern (Cook & Cook, 2017; Shadish, Cook, & Campbell, 2002). Recall the *p* value tells us how likely the results obtained in a study would have occurred if the null hypothesis is true. This statistic has no relation to the procedures for choosing participants, assigning participants to a condition, or whether important characteristics of the participants are consistent with the broader population. Thus, the *p* value cannot be relied on to inform decisions about whether study findings (e.g., an intervention causing improved oral reading fluency) can be expected to produce similar results at the population or individual levels.

EXAMPLES FROM THE LITERATURE

To illustrate how *p* statistics are used and what they mean, we briefly review two of the many recent studies in the field of special education that reported *p* values.

Training Preservice Teachers

Sayeski et al. (2015) examined the effects of multimedia training modules on preservice teacher (i.e., college student) knowledge and skills related to phonology and phonics. At the end of the introduction section of their report, the authors posed multiple research questions, including "To what extent can a series of interactive, multimedia modules support participants' development of basic language knowledge and related literacy skills?" (p. 242). Although the authors did not state a null or alternative hypothesis, the research question can be used to derive hypotheses. For example, we can infer that the null hypothesis would be *The interactive multimedia modules will have no effect on basic language knowledge and related literacy skills at the end of the study*. Similarly, the alternative hypothesis would be *The interactive multimedia modules will have an effect on basic language knowledge and related literacy skills at the end of the study*.

The 76 study participants were randomly assigned to treatment and control conditions. The authors examined the groups before the study to ensure they were not different in ways that might explain differences at post-test (i.e., both groups had participants with similar literacy courses, background experiences, levels of interest in the course, and levels of motivation). The authors concluded "given the similarities across groups in terms of both participant characteristics and pre-intervention questionnaire items, it was not anticipated that differences in groups would account for differences in study outcomes" (Sayeski et al., 2015, p. 243). To evaluate the effects of the intervention, the researchers administered the Survey of Basic Language Constructs (Binks-Cantrell, Joshi, & Washburn, 2012), which evaluated phonological and decoding knowledge of the participants prior to and following the study.

The researchers found differences between the intervention and control groups at the end of the study. Specifically,

they found the intervention group ($n = 38$) had an average score of 24.47 after the study, compared to the control group's ($n = 38$) average score of 19.74. This difference was statistically significant ($p < .001$). In other words, there is less than a 0.1 percent probability these results occurred with the null hypothesis being true. Consequently, the authors concluded the results were statistically significant. Because statistical significance does not indicate practical significance, the authors also calculated and reported an effect size of $d = 0.91$, which they described as a large effect. Because of the low p value and the ruling out of possible explanations for the differences between groups (e.g., groups were similar, nobody withdrew from the study), the study provides initial support for the alternative hypothesis. The authors appropriately remarked that multimedia modules have the potential for improving literacy knowledge and skills of preservice professionals, and emphasized the importance of conducting additional research to clarify how multimedia modules can be useful.

A Multicomponent Reading Comprehension Intervention

Solis, Vaughn, and Scammacca (2015) conducted a randomized controlled trial to investigate the effects of a multicomponent intervention—involving vocabulary instruction, text-based instruction, grammar print structures, complex language structures, inference reading drills, and curriculum-based measurement—on the reading comprehension of at-risk ninth graders. Like Sayeski et al. (2015), Solis et al. posed research questions but did not specify null or alternative hypotheses. The first research question posed was, "When differences in verbal ability are controlled for, to what extent does a multicomponent reading intervention with adolescent students who are low in reading comprehension impact reading comprehension outcomes?" (p. 105). We can infer that the null hypothesis would be *The multicomponent reading intervention does not impact reading comprehension outcomes for adolescent students with low reading comprehension*, whereas the alternative hypothesis would be *The multicomponent reading intervention positively impacts reading comprehension outcomes for adolescent students low in reading comprehension*.

The study involved 44 ninth-grade students who were previously enrolled in a reading intervention class due to reading comprehension difficulties. Students were randomly assigned to either a treatment or business-as-usual comparison group. The intervention was delivered across the school year as an elective class that met for 90 minutes, two to three times a week. The researchers assessed student performance on three different tests of reading comprehension before and after delivering the intervention. For the sake of brevity, we focus on results related to the Woodcock-Johnson-III Passage Comprehension Subtest (WJIII-PC; Woodcock, McGrew, & Mather, 2007; results were relatively consistent across all three tests). After controlling for verbal ability, as measured by a standardized test, scores on the WJIII-PC rose from 77.4 at pretest to 81.8 at post-test for the treatment group, but fell very slightly (from 77.6 to 76.7) for the comparison group. Although the authors interpreted the effect size ($\eta^2 = 0.03$) as medium, the difference was not statistically significant ($p = .26$). In other words, even though the treatment group outperformed the comparison group, there is a 26 percent probability that the null hypothesis is true given these results.

Accordingly, the researchers did not reject the null hypothesis, and the alternative hypothesis was not supported. The authors discussed potential reasons why the intervention did not yield a statistically significant difference between the groups, such as the possibility that remediation of reading comprehension for secondary students with low verbal ability may require a more intensive intervention over longer periods of time (e.g., multiple years). Given the medium effect size and the relatively small number of participants, the authors noted that future studies involving larger samples of participants might detect statistically significant effects for the intervention.

TAKE-HOME MESSAGE AND CONCLUSION

Professionals who intend to remain current with professional practice and deliver evidence-based interventions often turn to peer-reviewed research in reputable journals for guidance. The p statistic associated with NHST is a ubiquitous though commonly misunderstood aspect of quantitative research that can perplex and mislead education professionals and researchers alike when reading and interpreting research studies. Our take-home message is that the p value only tells us how likely it is that the results from a study occurred if the null hypothesis is true. Given its narrow utility and the widespread misunderstandings about p values, professionals should carefully consider many aspects of the study (e.g., effect size, sample size, representativeness of sample, plausibility of alternative explanations for findings) in conjunction with the p value when interpreting and applying study findings. Importantly, the p value signals whether a result can be considered interesting and worthy of further investigation, but cannot—in isolation—tell us whether the null hypothesis should be rejected, findings are practically valuable, the alternative hypothesis is true, or results are generalizable.

We have provided a brief overview of the p statistic by describing its meaning in educational research, and have also articulated some common misconceptions to which researchers and professionals alike may subscribe. The tendency to misattribute meaning to the p value suggests that special educators should be very cautious about ascribing value to research results based solely on the p value. Determining the validity and practical value of a study depends on understanding not just whether a low p value was obtained, but whether the effect size was practically significant, the participants represent a population of concern, other studies have obtained similar results, and the intervention likely caused the observed differences. If professionals cautiously approach research findings, understand the experimental method used, and critically evaluate the results—including but not limited to the p value—they may better interpret and apply research

findings in order to improve the learning outcomes of exceptional students.

REFERENCES

Binks-Cantrell, E., Joshi, R. M., & Washburn, E. K. (2012). Validation of an instrument for assessing teacher knowledge of basic language constructs of literacy. *Annals of Dyslexia*, *62*, 153–171. https://doi.org/10.1007/s11881-012-0070-8.

Cook, B. G., & Cook, L. (2016). Research designs and special education research: Different designs address different questions. *Learning Disabilities Research & Practice*, *31*, 190–198. https://doi.org/10.1111/ldrp.12110.

Cook, B. G., & Cook, L. (2017). Sampling and special education research: Examining whether and how study results apply to you. *Learning Disabilities Research & Practice*, *32*, 78–84. https://doi.org/10.1111/ldrp.12132.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 699–725.

Goodman, S. N. (1993). P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, *137*, 485–496.

Goodman, S. (2008, July). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*, 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., et al. (2016). Statistical tests, P. *European Journal of Epidemiology*, *31*, 337–350. https://doi.org/10.1007/s10654-016-0149-3.

Hubbard, R., & Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*, 69–88. https://doi.org/10.1177/0959354307086923.

Nuzz, R. L. (2015). The inverse fallacy and interpreting p-values. *PMR: The Journal of Injury, Function, and Rehabilitation*, *7*, 311–314.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchins and Company.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284. https://doi.org/10.1037/0003-066X.44.10.1276.

Sayeski, K. L., Kennedy, M. J., de Irala, S., Clinton, E., Hamel, M., & Thomas, K. (2015). The efficacy of multimedia modules for teaching basic literacy-related concepts. *Exceptionality*, *23*, 237–257. https://doi.org/10.1080/09362835.2015.1064414.

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

Solis, M., Vaughn, S., & Scammacca, N. (2015). The effects of an intensive reading intervention for ninth graders with very low reading comprehension. *Learning Disabilities Research & Practice*, *30*, 104–113.

Sullivan, G. M., & Fein, R. (2012). Using effect size - or why the P value is not enough. *Journal of Graduate Medical Education*, *4*, 279–282.

Travers, J. C. (2017). Evaluating claims to avoid pseudoscientific and unproven practices in special education. *Intervention in School and Clinic*, *52*, 195–203. https://doi.org/10.1177/1053451216659466.

Travers, J. C., Cook, B., Therrien, W. J., & Coyne, M. (2016). Replication research and special education. *Remedial and Special Education*, *37*, 195–204. https://doi.org/10.1177/0741932516648462.

Wasserstein, R.L., & Lazar, N.A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133. https://doi.org/10.1080/00031305.2016.1154108.

Woodcock, R. W., McGrew, K., & Mather, N. (2007). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.

About the Authors

Jason C. Travers is an associate professor and behavior analyst in the Department of Special Education at the University of Kansas where he coordinates the autism programs. He earned his doctorate at the University of Nevada Las Vegas and is a former public school special educator for learners with autism. He is interested in evidence-based practices, sexuality education, and technology-based interventions and supports for learners with autism.

Bryan G. Cook Professor of Special Education at the University of Hawaii, earned his PhD from the University of California at Santa Barbara. He is Past President of CEC's Division for Research, chairs the Research Committee for CEC's Division for Learning Disabilities, and co-edits Behavioral Disorders (the research journal of CEC's Council for Children with Behavioral Disorders). He is interested in evidence-based practices, bridging the research-to-practice gap, and examining the special education research base.

Lysandra Cook Associate Professor of Special Education at the University of Hawaii, earned her PhD from Kent State University. She is the coordinator for Project Laulima, a federal grant supporting the formation of a fully merged, co-taught teacher preparation program in elementary general and special education. Her scholarly interests include teacher preparation and evidence-based practices.