JSLHR

Review Article

Essential Statistical Concepts for Research in Speech, Language, and Hearing Sciences

Jacob J. Oleson,[a] Grant D. Brown,[a] and Ryan McCreery[b]

Purpose: Clinicians depend on the accuracy of research in the speech, language, and hearing sciences to improve assessment and treatment of patients with communication disorders. Although this work has contributed to great advances in clinical care, common statistical misconceptions remain, which deserve closer inspection in the field. Challenges in applying and interpreting traditional statistical methods with behavioral data from humans have led to difficulties with replication and reproducibility in other allied scientific fields, including psychology and medicine. The importance of research in our fields of study for advancing science and clinical care for our patients means that the choices of statistical methods can have far-reaching, real-world implications.
Method: The goal of this article is to provide an overview of fundamental statistical concepts and methods that are used in the speech, language, and hearing sciences.
Results: We reintroduce basic statistical terms such as the *p* value and effect size, as well as recommended procedures for model selection and multiple comparisons.
Conclusions: Research in the speech, language, and hearing sciences can have a profound positive impact on the lives of individuals with communication disorders, but the validity of scientific findings in our fields is enhanced when data are analyzed using sound statistical methods. Misunderstanding or misinterpretation of basic statistical principles may erode public trust in research findings. Recommendations for practices that can help minimize the likelihood of errors in statistical inference are provided.
Supplemental Material: https://doi.org/10.23641/asha.7849223

Statistical techniques are not magical or mysterious; rather, they are tools designed to quantify scientific evidence in various ways. Each method is built upon a mathematical foundation and has well-defined appropriate uses and requirements. Understanding more about these foundations, as well as the assumptions made by statistical procedures, can help investigators to adopt the most appropriate statistical method for the problem at hand, leading to more reliable and replicable results. Most traditional statistical methods follow the frequentist philosophy, in which models are fit via maximum likelihood or ordinary least squares. There are, of course, alternate perspectives on statistical inference, including Bayesian statistics and algorithmic modeling/machine learning. In addition, there are many techniques in nonparametric inference and variations on likelihoods (e.g., partial likelihood, semipartial likelihood, quasilikelihood, pseudolikelihood). Our focus in this article, however, is primarily on frequentist hypothesis testing using maximum likelihood, which is at the core of most applied science, as well as introductory statistics curricula. We expect that readers of this article have had at least one basic course in statistics. Researchers in speech, language, and hearing sciences are used to designing studies to learn about specific topics of interest, and many do perform their own statistical analyses to answer their hypothesis questions. Therefore, our goal is not to review all of the basics of statistics; rather, our goal is to highlight common errors and misconceptions in statistical approaches that can help scientists to avoid common statistical pitfalls in their research. In a companion article, some advanced methods for analyzing data in speech, language, and hearing sciences will be highlighted (Oleson, Brown, & McCreery, 2019).

[a]Department of Biostatistics, University of Iowa, Iowa City
[b]Boys Town National Research Hospital, Omaha, NE
Correspondence to Jacob J. Oleson: jacob-oleson@uiowa.edu

Editor-in-Chief: Katherine Gordon
Editor: Frederick Gallun
Received June 15, 2018
Revision received September 17, 2018
Accepted November 20, 2018
https://doi.org/10.1044/2018_JSLHR-S-ASTM-18-0239
Publisher Note: This article is part of the Research Forum: Advancing Statistical Methods in Speech, Language, and Hearing Sciences.

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

The use of statistical methods with underlying assumptions that do not match the data can have potentially serious consequences for the accuracy and reproducibility of scientific results. A recent analysis in major behavioral psychology journals indicated that approximately half of articles published between 1985 and 2013 contained at least one statistical error, and around 12% of published articles contained a statistical error that would have altered key findings of the study (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Widespread findings of errors in statistical reporting and interpretation are believed to have contributed to an inability to replicate key scientific findings from the literature in psychology (Pashler & Wagenmakers, 2012) and medicine (J. P. Loannidis, 2005). Problems with replication of statistical results erode public trust in science and can reduce the impact of scientific findings. Fortunately, many current problems related to the lack of transparency and reproducibility in scientific research can be resolved through increasing statistical proficiency of scientists and promoting open and transparent practices in the sharing of the code used for statistical analyses and data (Peng, 2015). To be clear, clinical and scientific experts do not need to become statistical experts, but they should recognize the statistical principles involved in their study design and their statistical analysis plan. Moreover, they should seek out collaborations with statistical experts to be involved with the study design and the statistical analyses. Study teams should include statistical expertise early in the development process to help design the study, to set up a clear and appropriate analysis plan, and to ensure appropriately analyzed and presented results.

Consider a research study in which we want to compare differences in the mean speech perception for children with hearing loss to children without hearing loss. There are multiple statistical approaches that could be applied to analyze differences between groups. Although scientists in nearly every discipline are familiar with the use of a two-sample Student's *t* test in this situation, confusion may arise when the assumptions of independence, normality, and equal variances are violated. An investigator must decide whether to immediately jump to a nonparametric analysis, implement a Welch adjustment for unequal variances, employ a repeated-measures analysis of variance (ANOVA), or transform the outcome variable, among many other options. All of these possible alternatives may produce the same conclusion, meaning the statistical approach we choose is inconsequential in many cases. On the other hand, we might obtain wildly different results depending on our choice of statistical method; understanding why the violation of assumptions can alter the behavior of statistical procedures is critically important to good statistical practice and making inferences based on the results of statistical analyses.

In most cases, there is not a single statistical approach that must be applied to solve a given research question, even when there is a clear tradition or common practice. Importantly, every statistical method comes with advantages and limitations. Practitioners should be familiar with the benefits and drawbacks of the procedures they employ and use that knowledge to choose methods that credibly answer the research question of interest. Our goal is not to provide detailed mathematical explanations of these tradeoffs but rather to focus on intuitive and practical recommendations when performing analyses, reporting results, and interpreting findings. We begin by discussing significance reporting and introducing the basic ideas behind hypothesis testing. We next discuss the ubiquitous use of the *p* value and what every user of a *p* value should know. The discussion of statistical significance via the *p* value is followed by measures of clinical significance through the appropriate use of effect sizes. We conclude by offering thoughts on regression analysis, model selection, and multiple comparisons.

Statistical and Clinical Significance

Significance Reporting

Research findings are often reported in terms of statistical and clinical significance, but it is very easy for both the producers and consumers of research to misuse or misinterpret these tools. Statistical and clinical (or practical) significance provide different pieces of important information. Ideally, both statistical and clinical significance should be described in reports and analytical results. Recall that statistical tests are generally devised as a choice between the null hypothesis (e.g., the average speech perception score for children with hearing loss is the same as the average speech perception score for children with normal hearing) and an alternative hypothesis (e.g., the average speech perception scores differ between children with and without hearing loss). To test a hypothesis is to ask whether we have enough scientific evidence to "reject the null hypothesis" and conclude that there is enough scientific evidence to conclude that the alternative hypothesis must be true (e.g., the speech perception means for children with hearing impairment and children with normal hearing are different). When we "do not reject the null hypothesis," we write that there is not enough evidence to conclude that the group means differ; we do not make positive conclusions (e.g., "the two group means are actually equal"). The phrase "statistical significance" simply indicates that we have satisfied a prespecified rule (how far apart the means are relative to the standard error), which allows us to reject the null hypothesis in favor of the alternative hypothesis. This is most commonly assessed using *p* values, which we discuss in more detail in the *p* Values section. The result of our reject or do-not-reject decision can be either correct or incorrect. If we reject the null hypothesis in favor of the alternative hypothesis, but the null hypothesis is actually true, then we have made a Type I error. For example, a Type I error occurs when we claim two groups have different means when the means are in fact equal. If we do not reject the null hypothesis, but the alternative hypothesis is the true one, then we have made a Type II error. This transpires when we do not claim that two group means are different when they really are different.

Clinicians and researchers should recognize that statistically significant findings are not always practically or

clinically significant. Statistical and clinical significance are often decoupled for large data sets because *p* values are strongly affected by the sample size. For example, we may find that the mean ages between two participant groups are significantly different statistically but that the observed mean difference is only 0.08 years (1 month). Unless this difference occurs at a point in development where the outcome of interest is likely to change rapidly over a short period, an age difference of 1 month between groups is unlikely to be clinically important. Reporting just the *p* value without highlighting the practical implications of the 1-month difference could be misleading.

Confidence intervals can help to contextualize significance tests and effects. Confidence intervals give a range of plausible values for a particular effect of interest. As such, they can be used to test hypotheses by checking whether the lower and upper bounds contain the hypothesized value of the quantity of interest. In addition, confidence intervals can also give an impression of clinical significance, which is discussed in more detail below in the Effect Sizes section in the context of "effect size." Briefly, an effect size can be as simple as reporting parameter estimates such as the sample mean, the mean difference between two groups, or the slope estimate in a regression analysis. Confidence intervals are subject to the same Type I and Type II error rates as significance testing more broadly, and the specific intervals given are often misinterpreted by researchers. Nevertheless, confidence intervals provide additional clinical effect information beyond the simple statistical significance of a finding and are discussed in more detail in the Effect Sizes section.

Unlike *p* values, the practical size of an effect does not change based on a sample size and provides a crucial piece of information that should be reported. In general, more time and space in scientific reports should be devoted to the discussion of the practical significance and scientific impact of detected effects. To be most impactful, scientific publications should convey both the evidence of an effect (statistical significance) and the practical impact of the detected effect (clinical significance). In addition, because different scientific audiences may have different tolerances for Type I errors, we recommend that researchers report actual *p* values, rather than just whether or not an effect is significant. Whereas preference or convention leads many researchers to be satisfied with a 5% Type I error rate ($p < .05$), others may prefer 1% or 0.1%, especially if the consequences of such an error would have a major impact on the field or alter an established scientific premise or clinical practice. In the next two subsections, we discuss *p* values and effect sizes in greater detail.

p Values

Although every introductory statistics class requires students to memorize their definition, *p* values have long been misunderstood and are often misused, misinterpreted, and increasingly criticized. At least one journal (Basic and Applied Social Psychology) has even taken the step of banning *p* values (Trafimow & Marks, 2015), and another has suggested using a more conservative *p*-value threshold to determine statistical significance to .005 (J. P. A. Loannidis, 2018). These controversies have led to ongoing debate and discussion across a wide range of scientific fields about the appropriateness of *p* values as the main criterion for judgments about statistical significance. To shed more light on the proper use and interpretation of *p* values, the American Statistical Association released a statement in 2016 (Wasserstein & Lazar, 2016) with six principles to consider for the proper use and interpretation of the *p* value. They are the following:

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 studied hypothesis 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.*

As we tackle the question of the usefulness of the *p* value, we need to understand what it is and what it is not. By definition, the *p* value is the probability of observing a test statistic, which is as extreme as or more extreme than that which was observed, assuming that the null hypothesis is true. If this probability is sufficiently low (e.g., below an acceptable Type I error rate) for the purposes of a given study, then we consider this sufficient evidence that the null hypothesis is likely not true and that we may favor the alternative hypothesis. The *p* value is not a tool for deciding which of two competing hypotheses is more likely, given the observed data.

In the frequentist statistical paradigm (the most common approach to statistics and data analysis), the hypothesis is not random, so we do not assign a measure of probability directly to it. From this perspective, the actual hypothesis is either true or false, and this fact does not change simply due to our inability to know the truth with certainty (e.g., the two means are equal, or they are not equal). Instead, we measure the likelihood that the data could have arisen if the null hypothesis were true. In order to accomplish this, we assume that the null hypothesis is true (e.g., the mean speech perception rates are equal for children with hearing impairment and children with normal hearing) and then define a rule to reject the said hypothesis if the observed data would be sufficiently improbable if that hypothesis were indeed true (e.g., these means are more than 2 *SD*s apart). If the observed data would have been unlikely to have arisen under the conditions specified in the null hypothesis (e.g., the sample means for our two groups are too far apart for them to have come from two distributions with the same mean), we consider that evidence against the null hypothesis.

This process does not allow for a probability comparing the two competing hypotheses.

In some settings, researchers do desire a method to compare the likelihood of competing hypotheses. In such settings, Bayesian statistical approaches may be more useful. In the Bayesian paradigm, one can evaluate the probability of the null hypothesis being true versus the probability of the alternative hypothesis being true, because the Bayesian interpretation of probability is fundamentally different from the frequentist interpretation. Broadly speaking, Bayesian probability is a tool to quantify knowledge and uncertainty. Bayesian statistics uses probability to describe what is known before data are collected and to update that knowledge based on how much is learned after collecting data. More detail on Bayesian statistics is given in this issue in Oleson et al. (2019) and McMillan and Cannon (2019).

p values suffer from other drawbacks besides the confusing definition. Consider a two-sample problem to compare two group means such as testing whether the mean speech perception score of the population that is hard of hearing is significantly different from the mean speech perception score of the population with normal hearing. Although the null assumption that population means are equal is statistically useful, in real-world settings where we cannot randomly assign group membership, we know a priori that the two population group means are not the exact same value. In our example of speech perception scores, it seems improbable that a language outcome score for the group with hearing impairment has the same population mean as the group with normal hearing. Our primary interest is to determine whether a statistically detectable and clinically relevant difference exists. Even in experimental settings, if there is a difference worth testing, there is often a good reason to suspect that at least an infinitesimal difference between group means exists. Nevertheless, we assume that the population means are equal in a null hypothesis. In order to demonstrate evidence of a difference between them, we must first naively assume that there is no difference. Whenever the population means are even slightly different, we know that we can make a *p* value small enough to declare statistical significance simply by choosing a large enough sample size (e.g., a high enough power). Indeed, researchers who fail to reject null hypotheses are often advised to increase their sample size. This practice is not statistically appropriate in part because it does inflate the Type I error. In this sense, *p* values by themselves often tell us more about sample size than anything practically meaningful about the size of the effect. Moreover, *p* values are constructed under the statistical model that we assumed: Any departures from the assumptions of that model impact the reliability of the *p* value.

If we use common, incorrect statistical practices in speech, language, and hearing science research, including sampling from finite populations without correction, obtaining correlated samples that are not accounted for, or sampling from populations with different distributions from those assumed by particular statistical procedures, the applicability and reliability of *p* values are greatly diminished. Therefore, the statistical analysis that is used should be carefully determined to appropriately reflect the problem at hand and that the relevant assumptions should always be considered and checked prior to making inferences based on statistical results.

Effect Sizes

An "effect" or "effect size" is a measurement of a phenomenon of interest, whether an expected change in an outcome in response to a treatment, a difference between group means, or a ratio of the odds of an outcome in two groups. Importantly, the specific effect size that is chosen should provide insight concerning a meaningful study question.

Effect sizes can be subdivided into two groups: "standardized" or "relative" effect sizes and "unstandardized" or "absolute" effect sizes. Unfortunately, which kind of effect is being presented in a given analysis is not always clear. Cohen's *d* is a commonly used relative effect size, whereas our previous example of 1-month difference in age is an example of an absolute effect size. Many formulas for the power of statistical procedures and sample size of study designs are given in terms of Cohen's *d*, because the power to reject a null hypothesis depends on the relative extremeness of the alternative hypothesis; Cohen's *d* gives a standardized measure of how far an estimated effect is from a hypothesized value, relative to its standard error. This gives a convenient scale on which to compare and interpret mean differences but suffers from the same problem as *p* values: Cohen's *d* does not measure the practical significance of a difference but is instead more closely related to statistical significance. Using our previous example, if we detect a statistically significant difference in mean age between two groups, we can equivalently expect that the difference in means is large relative to its standard error. Lenth (2012) provides an excellent summary of the shortcomings of comparisons of relative effect sizes. A clinically meaningful effect is typically better evaluated by examining the magnitude of the absolute mean difference.

Absolute effect sizes for hypothesis tests comparing means are simply the difference between the means of interest. Statistical software for most statistical modeling frameworks will output absolute effect sizes based simply on the parameters reported by statistical software. Obtaining relative effect sizes in the style of Cohen's *d* in modern statistical models can be somewhat more complex given the potential difficulty of estimating the standard error of the difference between groups, which is the denominator term in the calculation of Cohen's *d*. Although methods exist to compute relative effect sizes for more complicated models (Brysbaert & Stevens, 2018; Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012; Westfall, Kenny, & Judd, 2014), we generally prefer to rely on more interpretable absolute effect sizes.

Confidence intervals provide another perspective on absolute effect sizes. Although they suffer from some of the same interpretability challenges as the *p* value, they do give us more direct information on the range of plausible values for the parameter of interest. Consider the difference between speech perception scores between two groups again. For two independent groups, a two-sample independent

groups confidence interval is typically computed using the difference in the means and a weighted average of the variances of the two groups (a pooled variance). Note that we say typically because we are assuming equal variances between the two groups and that the difference in means approximately follows a normal distribution, and any deviations from those assumptions would require alternate methods. The confidence interval is then centered at the mean difference, and we add and subtract the standard error (square root of the pooled variance) times our critical value. The result is an upper limit and a lower limit of confidence of where we believe the true population mean difference might be. The interpretation challenges come from the fact that the performance guarantees (confidence levels) concern the repeated process of how confidence intervals are constructed in general, rather than the specific numbers produced in one single statistical analysis. For example, in a study where the mean difference in language standard scores between two groups of children is 4.51, with a valid 95% confidence interval of [3.30, 5.72], we have confidence that intervals constructed in this way will contain the true language standard score 95% of the time. No such statement can be made about the specific interval of 3.30–5.72. Even so, this range does give plausible values that are consistent with the observed data—a valuable addition to the practical understanding of the absolute size of an effect.

When reporting study results, it is important to report measures representing both the statistical significance and clinical significance of the study. Absolute effect sizes generally carry the most information regarding clinical importance for the study and should be reported. Generally, this should be in terms of the parameter estimate and the confidence interval for the estimate. Relative effect sizes can also be informative in some situations but carry additional complexity and can conflate statistical and practical significance for practitioners and readers.

Regression Topics

Regression and ANOVA

A common statistical technique employed in the field is regression analysis. In its most basic form, a regression analysis includes one dependent variable that is related to the outcome via a line, although regression models can be made much more complex. The goal is generally to test whether all or some of the independent/explanatory variables in the model are related to the dependent/outcome variable. Those independent variables may be continuous or categorical. In the event that there is a single categorical independent variable, then the resulting regression model is commonly referred to as *ANOVA*. Although classical regression and ANOVA were developed separately, ANOVA is simply a special case of regression, where regression can accommodate more complicated relationships between the dependent variable and the independent variables.

In a regression-type model, the slope parameter estimates themselves provide a measure of effect size. In cases where multiple groups are present (such as in ANOVA or regression with categorical variables), the effect of interest may be given by the estimate of a contrast or a single parameter estimate comparing one group to a reference group. See Bring (1994) for a full discussion of regression effect sizes. In Walker et al. (2014), regression models were used to investigate relationships among predictor variables and service delivery for a group of children who were hard of hearing. The independent variables were gender, test site, maternal education level, immediate family history of hearing loss, and degree of hearing loss measured by better-ear pure-tone average. They found that, after controlling for all of the variables, only degree of hearing loss was significantly related to the dependent variables age at first diagnostic evaluation ($\beta = -0.36$, $p = .003$), age at hearing loss confirmation ($\beta = -0.42$, $p = .001$), and age at hearing aid fitting ($\beta = -0.37$, $p = .011$). By reporting the β (slope) values, we can immediately determine how better-ear pure-tone average is related to each of these age-related outcome variables. They also found that only gender was significantly related to length of the delay between hearing loss confirmation and enrolling in early intervention ($\beta = -3.34$, $p = .024$). Through reporting of the β value as an effect size, we immediately know that girls had an estimated delay that was 3.34 months shorter than boys, and we can readily assess the clinical impact those months will have on the children.

In ANOVA, we are often confronted with the need to adjust our outcome variable of interest by a covariate that is continuous. A classical approach to performing this adjustment is analysis of covariance (ANCOVA). Recall that ANOVA is designed to compare group means by measuring the between-group variance relative to the within-group variance. ANCOVA is a method that mathematically adjusts the group means to take into account the values of the covariate (e.g., age) and then performs ANOVA on the covariate-adjusted means. However, just as ANOVA is equivalent to a regression analysis that includes only a categorical variable, ANCOVA simply adds a continuous variable to be adjusted for to that regression model. For example, an ANCOVA that examines differences in speech recognition for children who are hard of hearing and children with normal hearing adjusting for age is the same as a linear regression analysis that includes age and hearing status (normal hearing or hard of hearing) as predictor variables. General linear hypothesis tests can still be performed within the regression model framework. General linear hypothesis tests include what many in the field call *post hoc tests.* Technically, *post hoc tests* are those tests, usually pairwise comparisons, that are only considered after the results of the global tests are found. If these pairwise comparisons are planned from the beginning of the study, then they are not technically post hoc tests.

Moreover, regression models and their generalizations offer a great deal of flexibility to include additional covariates, add random effects, modify the residual error structure, do nonlinear transformations, and more. The addition of random effects to a regression model to create subject-specific curves and account for within-subject correlation is

referred to as linear mixed-effects regression models or linear mixed models. More details on mixed models can be read in our companion article (Oleson et al., 2019). Although ANOVA and ANCOVA are familiar statistical procedures with a lot of history, it is preferable to use the regression framework for the task of comparing group means.

Model Selection

Often, we encounter the situation of deciding what independent variables to include in our regression model. This process of deciding which variables to include, and which to not include, is known as *model selection*. For example, Walker et al. (2014) were interested in predictors of the dependent variable age at hearing aid fitting. The independent variables were gender, site at testing, maternal education level, immediate family history of hearing loss, and degree of hearing loss. The variables included in the final model were included because of specific hypotheses about the factors that influence hearing aid use from theory and the previous literature on hearing aid use in children. Model selection refers to deciding what subset of the full list of independent variables should be included in the final model.

Model selection is broader than just regression and arises in all forms of statistical inference. Investigators must identify a model that can address the research questions of interest using the variables that were collected as data. Ideally, scientific theory should be the foundation for the process of model selection, comparing among a small set of scientifically plausible models. Some approaches to model selection are based on the principle of parsimony: The best statistical model is that which includes all the essential variables and nothing more. In practice, however, the most parsimonious model for a specific research question can be difficult to identify. In many exploratory studies, we gather data without a full understanding of what the important relationships are, which covariates need to be adjusted for, and what patterns of correlation in longitudinal studies are likely to be appropriate. Although model selection is a large topic area and an active field of statistical research, common techniques in a frequentist setting can be broadly divided into two categories: algorithmic approaches and researcher-driven exploration. The process of model selection is crucially important and can impact the outcome and reliability of hypothesis testing procedures, as well as the inferences and conclusions of a research study.

At the core of any model selection procedure is the ability to compare selected models of interest to determine which better fits the data. This comparison can be accomplished for nested models in a regression setting using F tests, which test the null hypothesis that the collection of variables to be added to the model does not explain a significant amount of variability in the outcome above and beyond the currently included variables. More generally, we can compare frequentist models using information criteria such as the Akaike information criterion (AIC). Models with smaller AIC values are considered to have better fit to the data. Algorithmic model selection approaches build on these concepts by sequentially considering modifications to a model. For example, forward selection starts with a base model and proceeds to consider adding terms to a model. Terms are added at each step if they meet an inclusion threshold, which can be based, for example, on p values or information criteria. Similarly, backward selection starts with a large model and considers terms for removal. Researcher-driven exploration may use some of the same tools to compare models, but decisions on what models to explore and select are made by the researcher. This process of model selection may be informed by scientific knowledge, formal model comparisons, or conscious/subconscious preference for certain model features or results.

Although it is clear why informal exploratory analysis can lead to problems with inflated Type I error rates and reproducibility, many practitioners are surprised to discover that formal techniques such as stepwise model selection are similarly problematic. In each case, the output generated by statistical software after a model selection routine is performed is indistinguishable from output obtained from a prespecified analysis. For most statistical procedures, the reported results are based on the assumption that data were collected according to a simple random sample from a large population, and then a single model was fit. These assumptions are not met when data-driven model decisions are made, and these assumption violations can endanger the validity of statistical inference.

To illustrate this point, we highlight several simulation results from a regression example. Replicate data sets (50,000) are simulated with 10 hypothetical covariates and a normally distributed outcome measure and are generated under the global null hypotheses that all of the regression coefficients are zero. Each data set has $n = 100$ observations. Two primary hypothesis tests are of interest. First, we consider the setting where one covariate addresses a primary study question, and the other nine covariates are included to adjust for potential confounding. In the second, we consider the overall F test for the (true) null hypothesis that all of the covariate beta coefficients equal zero. For prespecified analyses conducted without model selection, we have theoretical guarantees that the Type I error rate will not exceed whatever nominal threshold we specify.

For a chosen Type I error rate of 0.05, we observe that choosing large or small models presents no problems in the absence of model selection. In line with our theoretical expectations, the observed Type I error rate for the analysis of just the single variable of interest produced a Type I error rate of 0.0502, whereas an analysis of all 10 covariates with pairwise interactions produced a comparable Type I error rate for this comparison of 0.0496. The overall F-test results were similar.

In the presence of forward stepwise model selection based on AIC, however, the hypothesis test concerning our primary variable of interest had a Type I error rate of 0.0645, and the overall F test had a Type I error rate of 0.4314. Therefore, just by applying basic model selection algorithms, we are increasing our probability of incorrectly rejecting

the null hypothesis. Although these results are quite troubling, especially for overall F tests, the situation in real analyses may be even worse, as actual data tend to exhibit correlation among the independent variables or various types of confounding, driving up variability of parameter estimates and the likelihood of incorrect conclusions. With this in mind, it is incumbent upon practitioners to clearly delineate between exploratory analyses and confirmatory analyses in scientific reports and to clearly describe any model selection procedures that were employed in a study. By default, p values only attain their advertised performance for prespecified models. Code and complete results of the simulation are provided in Supplemental Material S1.

Multiple Comparisons

Statistical models may require comparisons of several effects within an overall model. Traditionally, scientists have been trained to adjust their statistical assumptions to be more conservative to account for multiple comparisons to avoid the increasing risk of statistical errors as the number of comparisons increases (e.g., Aickin & Gensler, 1996). However, there are also costs to many approaches to accounting for multiple comparisons, including reducing statistical power. There are different schools of thought when it comes to multiple comparisons and many relevant summary articles. See overviews by Saville (1990), Bender and Lange (2001), or Cao and Zhang (2014) for more in-depth discussions of this issue. Rather than provide a review of the myriad methods of accounting for multiple comparisons in statistical models, we lay out some general points to consider regarding the topic.

The primary reason that researchers are advised to do a multiple-comparisons adjustment is to strictly control the overall (familywise) Type I error rate. In summarizing the alternative approaches for adjusting for multiple comparisons, we will consider three different general approaches. We could (a) perform no adjustment and accept individual Type I error rates, (b) adjust our alpha level to preserve a predetermined familywise error rate, or (c) adjust the alpha level to allow for a specified acceptable false-positive error rate.

The first approach to multiple testing is applicable when an alpha level adjustment may not be required. Bender and Lange (2001) argue that multiple comparisons should be used in confirmatory studies for the primary outcome of interest and that they are not necessarily required for exploratory studies. The researcher should clearly define the primary outcome and identify which comparisons correspond to that outcome. This may be a small subset of analyses that are performed, where the rest of the analyses are secondary and perhaps dependent upon the results of the primary question. If we have a truly controlled and confirmatory analysis, then we do want to reasonably control this alpha level. There is also the question of making Type II errors where we do not detect an important difference. In speech, language, and hearing studies, where sample sizes tend to be small, the alpha level adjustments reduce the significance level, which also increases the Type II error. It is critical to report absolute effect sizes in these situations.

The most common approach implemented for multiple comparisons is to adjust the alpha level to preserve a predetermined familywise Type I error rate. The Bonferroni adjustment is the most common technique used. For all pairwise comparisons, some will choose a Tukey–Kramer adjustment, but there are many more ways of adjusting alpha to control familywise error rates. Although it is clear that these adjustments are conservative, it may not be as clear the assumptions that are being made by such an adjustment. In an adjustment to control the overall Type I error, a critical assumption is that all null hypotheses are correct and that we want to jointly make only a 5% chance of falsely rejecting at least one of those. However, we happen to know that it is highly unlikely that all null hypotheses are true, as we outlined in the Significance Reporting section. These adjustment techniques make the most sense in a highly controlled confirmatory study with a single outcome of interest. In many studies, especially those that include various covariates (e.g., age, hearing loss severity, gender), we do expect many of the null hypotheses to be false and would be surprised not to find significant evidence against them. For example, a study in which age is adjusted for as a known confounder is expected to find a significant effect due to age. As an alternative example, a study with a placebo arm, a known effective treatment arm, and a novel treatment would be expected to reject the null hypothesis that the existing treatment and placebo produce the same mean. Assuming the global null hypothesis as a basis for Type 1 error rate correction is often unrealistic and unnecessarily conservative.

The third approach is to adjust the alpha level to allow for an acceptable error rate. This differs from approaches that attempt to control the overall Type I error rate because the goal is to control the proportion of our "significant" results that are incorrect, rather than the probability of making any such mistakes. The false discovery rate (FDR; Benjamini & Hochberg, 1995) and the procedures for controlling it were developed for this purpose. This method makes more intuitive sense for how we think about testing, but it does not guarantee an overall prespecified Type I error rate. FDR-adjusted p values control the number of false discoveries (Type I errors). For example, an FDR of 0.05 implies that, on average, 5% of significant tests will result in false positives under the null hypothesis. The FDR-adjusted p value can be far less conservative than a Bonferroni adjustment but still addresses the goals of controlling Type I errors when multiple comparisons are made.

Discussion

The goals of this article were to (a) present statistical concepts and methods that we regularly see implemented and misunderstood in speech, language, and hearing sciences research and (b) offer additional insights or alternatives to those concepts and methods. Like other disciplines,

the field of statistics has evolved over time to accommodate the realities of modern research involving human subjects. Although the traditional statistical methods that we have discussed in this article still have relevance and specific uses, often other tools can better and more accurately answer the research question of interest and offer greater flexibility for more complex research designs. It is up to scientists in the field to continue educating themselves in modern statistical practice to find statistical approaches that fit their specific goals and to work with statistical experts throughout the research process.

This education begins with a better understanding of the *p* value and its worth. When the *p* value is reported in conjunction with an appropriate measure of the effect size, then it gives the researcher important information about the study findings, both statistically and clinically. It is imperative to consider the information actually conveyed by a particular measure of effect size and how it informs the statistical results and practical implications of a given study. Although many of the concepts described in this review are basic statistical principles, the misunderstanding or misinterpretation of these concepts is a substantial threat to the validity of our research findings. To help minimize the potential for statistical errors, researchers in the speech, language, and hearing sciences can do the following:

1. Understand how *p* values are calculated and what *p* values represent. The 2016 recommendations of the American Statistical Association regarding *p* values (Wasserstein & Lazar, 2016) can help scientists to avoid common misconceptions about *p* values and understand the difference between statistical and clinical significance.
2. Report measures of statistical and clinical or practical significance as part of articles or reports summarizing research findings. Presented together, metrics of statistical and clinical significance enhance the interpretation of research and make the effects meaningful for clinicians and patients who are much more likely to be interested in the magnitude of the effect in terms of a specific outcome rather than the statistical significance alone.
3. Include confidence intervals for effect sizes. The appropriate use of confidence intervals conveys the potential range of plausible values around an effect and can allow consumers of research to understand the influence of variability on the precision of reported effect sizes.
4. Choose statistical approaches that allow for modeling complexity over methods with more rigid assumptions such as ANOVA or ANCOVA. Using flexible methods such as regression or mixed models can expand the breadth of research questions that can be evaluated beyond examining differences between groups or across conditions.
5. Develop model selection procedures based on scientific knowledge and theory that are parsimonious solutions to complex phenomena. Different iterations of the same statistical model should be compared using established information-based methods, such as comparisons using AIC or Bayesian information criterion.
6. Use methods for controlling for multiple comparisons within statistical models that are specific to the goals of the research rather than always using overly conservative approaches that control for familywise error rate, such as Bonferroni adjustment. Decisions about the appropriate method for controlling for multiple comparisons should occur prior to the statistical analysis based on the goals of the comparisons and design of the study.

The goal of reporting results from statistical analyses in articles should be to present new clinically relevant findings or suggest future research opportunities. Readers of the work should be able to replicate the experiment and the analysis. The statistical methods should be written with enough detail that a data analyst could read it and replicate the analysis. Furthermore, code and data should be provided when feasible to promote transparency and reproducibility. Statistical methods are designed to provide good results under uncertainty but always include the possibility of error. With this in mind, replication is an essential part of scientific progress, and wherever possible, researchers should facilitate these efforts. There are many opportunities to apply these principles in the speech, language, and hearing sciences, and it is our hope that this article will help orient investigators to the problem of selecting appropriate and robust statistical models.

Acknowledgments

This research was supported by grants from the National Institute on Deafness and Other Communication Disorders Grant R01 DC013591, awarded to Ryan McCreery. Additionally, we thank the editor and two anonymous referees for many helpful comments and suggestions on drafts of this article.

References

Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health, 86*(5), 726–728.

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—When and how. *Journal of Clinical Epidemiology, 54*(4), 343–349.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, Statistical Methodology, 57*(1), 289–300.

Bring, J. (1994). How to standardize regression coefficients. *The American Statistician, 48*(3), 209–213.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1), 9. https://doi.org/10.5334/joc.10

Cao, J., & Zhang, S. (2014). Multiple comparison procedures. *Journal of the American Medical Association, 312*(5), 543–544.

Lenth, R. V. (2012). Some practical guidelines for effective sample size determination. *The American Statistician, 55*(3), 187–193.

Loannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294*(2), 218–228.

Loannidis, J. P. A. (2018). The proposal to lower *p* value thresholds to .005. *Journal of the American Medical Association, 319*(14), 1429–1430.

McMillan, G. P., & Cannon, J. B. (2019). Bayesian applications in auditory research. *Journal of Speech, Language, and Hearing Research, 62,* 577–586. https://doi.org/10.1044/2018_JSLHR-H-ASTM-18-0228

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*(4), 1205–1226.

Oleson, J. J., Brown, G. D., & McCreery, R. (2019). The evolution of statistical methods in speech, language, and hearing sciences. *Journal of Speech, Language, and Hearing Research, 62,* 498–506. https://doi.org/10.1044/2018_JSLHR-H-ASTM-18-0378

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence. *Perspectives on Psychological Science, 7*(6), 528–530.

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance, 12*(3), 30–32.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *The American Statistician, 44*(2), 174–180.

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's f^2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology, 3,* 111.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1–2.

Walker, E. A., Holte, L., Spratford, M., Oleson, J., Welhaven, A., & Harrison, M. (2014). Timeliness of service delivery for children with later-identified mild-to-severe hearing loss. *American Journal of Audiology, 23*(1), 116–128.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045.