Perspective from the Literature on the Role of Expert Judgment in Scientific and Statistical Research and Practice

Naomi C. Brownstein*

Abstract

This article, produced as a result of the Symposium on Statistical Inference, is an introduction to the literature on the function of expertise, judgment, and choice in the practice of statistics and scientific research. In particular, expert judgment plays a critical role in conducting Frequentist hypothesis tests and Bayesian models, especially in selection of appropriate prior distributions for model parameters. The subtlety of interpreting results is also discussed. Finally, external recommendations are collected for how to more effectively encourage proper use of judgment in statistics. The paper synthesizes the literature for the purpose of creating a single reference and inciting more productive discussions on how to improve the future of statistics and science.

Keywords: Bayesian modeling, Hypothesis testing, Inference, Insight, Interpretation, Knowledge, Recommendations, Significance, Statisticians, Subjectivity

*This work was not supported by any grant. However, the author would like to thank Jeff Harman, Tom Louis, Tony O’Hagan, and Jane Pendergast for their helpful comments during the preparation of this manuscript.
1 Introduction

As a quantitative discipline, statistics is often considered by practitioners as objective in its methods, which proliferate throughout the scientific enterprise. Basic statistical methods, including Frequentist hypothesis tests and $p$-values, are ubiquitous in academic literature. The American Statistical Association (ASA) recently warned the research community about common misuses of $p$-values in a statement (Wasserstein & Lazar 2016) that has received widespread attention in a variety of fields, even outside of statistics. As a follow-up, the ASA organized the 2017 Symposium on Statistical Inference (SSI), where statisticians and consumers of data presented research, contemplated these issues, and brainstormed possible solutions. Yet, the statistical community remains divided on explicit recommendations for researchers to improve their quantitative practice (Matthews et al. 2017). Due to the inherently quantitative nature of the field of statistics, much of the conversation on the $p$-value statement has revolved around quantitative solutions. Discussions of the statistical properties of these and other remedies are included elsewhere in this special issue.

Despite perceived objectivity of the field, statisticians previously argued that uncertainty and choice abound in the scientific process, from the definitions of questions of interest to the analysis and interpretation of results (Gelman & Hennig 2017, Goldstein et al. 2006, Berger & Berry 1988). Hence, expert judgment is required to implement scientific research (Bertolaso & Sterpetti 2017). A session at SSI examined the role of judgment in statistical and scientific practice. Session participants present their joint expert opinions on the topic in Brownstein et al. (2018). While preparing the paper, a wealth of related literature was collected and synthesized. The present article provides an overview of the literature on the role of expert judgment in statistics and science.

First, section 2 defines expertise as it relates to this paper. Then, sections 3, 4 and 5 discuss the role of expert judgment in Frequentist hypothesis testing, Bayesian inference, and general interpretation. Recommendations are provided in section 6 for stakeholders in education, publishing, and funding. Finally, appendices are provided for interested readers including overviews of the Frequentist and Bayesian paradigms.
2 What is Expert Judgment?

Before discussing the role of expert judgment in science, definitions of the relevant terms are needed. Weinstein (1993) provides the following definitions of expert, expertise, and expert opinion that are invoked throughout this paper:

1. An individual is an expert in the ‘epistemic’ sense if and only if he or she is capable of offering strong justifications for a range of propositions in a domain.

2. An individual is an expert in the ‘performative’ sense if and only if he or she is able to perform a skill well.

3. A claim is an ‘expert opinion’ if and only if it is offered by an expert, the expert provides a strong justification for it, and the claim is in the domain of the expert’s expertise.

4. ‘Expertise’ is the capacity either to offer expert opinions or to demonstrate one or more skills in a domain, and expertise in a domain does not entail expertise in the entire range of the domain.

In brief, epistemic experts possess deep knowledge about a field and are considered credible by others (Carrier 2010), while performative experts are adept at completing actions for the field. The late Stephen Hawking (Gribbin & White 2016) is a famous epistemic expert in physics. An olympic athlete, such as Usain Bolt (Gómez et al. 2013), exemplifies performative expertise in their sport. The two types of expertise frequently overlap. Physicians, for instance, are both versed in fields such as anatomy and proficient at diagnosing ailments, performing medical procedures, and prescribing treatments and cures.

Expert judgment involves one or more evidence-based claims proposed by credible experts. Judgments may arise when multiple conclusions or actions are possible and supported by credible evidence. For example, multiple treatments may be possible for a condition, and two physicians may disagree on the preferred treatment plan for the same patient. Sections 2.1 and 2.2 relate these definitions to statistics and science.
2.1 Statistical expertise

Merriam-Webster Online Dictionary (2018b) defines statistics as “a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.” In general, like any professional skill, statistical analysis should be conducted by practitioners who are qualified based on their knowledge of and experience with statistical procedures and principles. Example key principles include knowledge of modeling assumptions and diagnostics, as well as experience using and interpreting statistical software.

Based on the definition of expert in Weinstein (1993), an active statistician, defined as “one versed in or engaged in compiling statistics” (Merriam-Webster Online Dictionary 2018a), is clearly considered an expert. Similar to the ability of a primary care doctor to identify and treat general health needs for their patients, statisticians are trained with knowledge and skill in choosing, creating, implementing, and validating statistical procedures for a wide variety of settings. Others can be considered statistical experts as well if they have sufficient relevant knowledge and training in statistical theory and practice. Both types of expertise are needed for valid statistics, even if divided among multiple people. A common model is for a person with extensive training in statistical theory to supervise the study design and statistical analysis, which is executed by an analyst facile in programming.

Statistics is a broad field with numerous subfields, rendering it impossible to know all aspects of each. Instead, analogous to how physicians choose medical specialties, most statisticians specialize in one or more subfields, such as genetics, Bayesian modeling, or survival analysis. Classifications of subfields in statistics are found in Schell (2010) and De Battisti et al. (2015). Specialized problems require expertise in different types of analytical methods. Identifying and deciding between reasonable analytical choices requires judgment, as described throughout this paper and highlighted by others (Francis 2017).

2.2 Content expertise

While statisticians are critical members of scientific teams, interdisciplinary collaborations include experts in multiple fields. For instance, to develop a clinical trial to test an oncology drug, expertise could be required in fields such as medicine, biology, and genetics. For the purpose of this article, a content expert, or subject matter expert, has expertise in a field
of science other than statistics. Content experts are key in advancing scientific endeavors, such as generating hypotheses based on knowledge of biological mechanisms, designing and supervising experiments using state of the art laboratory techniques, providing intuition on clinically meaningful values, and contextualizing statistical results. Communication of content knowledge to the statistician is also critical, as it insight may shape the study design and analysis in subtle ways, as described in (Brownstein et al. 2018).

3 Expert Judgment in Hypothesis Testing

This section discusses the judgment of experts as defined in Section 2 for Frequentist hypothesis testing. Topics include the role of expert judgment in hypothesis testing, including finding balance between between type I and II errors, multiple comparisons adjustment, and special considerations such as orphan drugs and non-inferiority trials.

3.1 Trade-off Between Type I and II Errors and Two Hypotheses

It is well known that for a fixed sample size, there is a trade-off between minimizing false positives (type I error) and maximizing true positives (power). Clearly, it is desirable both to have small error probabilities and high power. One way to minimize the probabilities of both errors simultaneously is to increase the sample size of an experiment (Asendorpf et al. 2013). However, this solution may not be feasible, as outlined in Section 6.3. Instead, experts should consider their preferred balance between the two types of errors and design their experiments appropriately. The statistician is responsible for engaging the subject matter experts in discussions to determine the levels of each type of error that are scientifically and ethically reasonable for their study. Current methods of using judgment to balance errors in Frequentist hypothesis testing are discussed along with their implications.

3.1.1 Minimizing Type I Error

For most experiments, the significance level is first set. In other words, the judgment is that type I errors should be set at a certain low level, after which other considerations can be made. The most famous value is 5%, usually chosen by convention. Rather,
Lakens et al. (2018) argues that researchers should determine their ideal level for their individual study beforehand using decision theory and report the decision and methodology in the manuscript. Unfortunately, this practice is uncommon; brief discussion is provided in section 3.1.3. Instead, most researchers compare to a single threshold, the customary value for which is currently under discussion. A summary of the discussion is provided.

Recent concern about the reproducibility in science focuses on the fact that interesting associations in the literature are not easily replicated in future studies (Ioannidis 2005). A hypothesized cause is that the observed type I error rate may far exceed that of the nominal significance level. This hypothesis inspired recent recommendations to drop the most commonly used significance level further, e.g. to 0.005 (Benjamin et al. 2017, Johnson et al. 2017, Johnson 2013). The goal is to minimize the probability of a type I error in hopes of filtering out weak effects and increasing the reproducibility of science as a whole.

On the other hand, there is concern that the reproducibility crisis may be due to a widespread lack of power due to inadequate power calculations (Marino 2017, Smaldino & McElreath 2016, Asendorpf et al. 2013). Calls to lower the type I error rate have been criticized for enabling underpowered studies if no other action is taken, such as raising funding levels to accommodate increased sample sizes (Asendorpf et al. 2013, Lakens et al. 2018). Further, Lakens et al. (2018) argue that the recommendation may be hurting the very cause that it champions by decreasing resources and incentives for replication studies.

3.1.2 Minimizing Type II Error

The paradigm of prioritizing a fixed significance level chosen by convention only may be inappropriate for specific applications, especially if excessive harm could be associated with a lack of discovery. For example, if the question evaluates the harm of a certain environmental factor, then the precautionary principle (Fischer & Ghelardi 2016, Fjelland 2016) states that it is better to err on the side of caution, as failing to detect underlying harm would result in failure to reduce personal harm for individuals currently affected and future individuals. In these cases, minimizing the type II error probability is judged to be of the utmost concern. Power calculations necessitate specification of particular realizations of the alternative hypothesis or effect size. Determination of clinically and statistically meaningful
effect sizes is highly dependent on the judgment of both statisticians and content experts; guidance is provided elsewhere (Murphy et al. 2014, Ellis 2010, Lakens 2013).

The story of Love Canal, a suburban neighborhood near Niagara Falls, prioritizes minimizing the harm of a type II error over that of a type I error. The general null hypothesis is the safety of the area; the alternative is an association between toxic chemicals from the Love Canal and increased cancer prevalence. The relative (bodily) harm of a type II error, declaring the area safe when it was hazardous to the residents, exceeded the (financial) harm of a type I error. Initially, an investigation erroneously concluded that the area was safe. The judgment of experts, namely a scientist with detailed observational data from the residents, helped diagnose the error in the initial investigation. Eventually, the area was closed. Briefly, the first study included a reasonable test to answer the wrong question. The original investigation tested the seemingly intuitive hypothesis that proximity to a chemical waste site was associated with more negative outcomes. The second investigation refined the hypothesis, namely that residents in homes near the site built over former stream-beds were more at risk than homes built over dry land. The latter test found a strong association based on sound biological principles with clear implications for evacuation prioritization. A detailed discussion of the story and the role of judgment is found in Fjelland (2016).

Similarly, prioritizing power, the United States Food and Drug Administration has a separate regulatory category for orphan drugs to allow new treatments to have a higher chance of adoption if no other therapeutic option is available to mitigate a rare condition (Braun et al. 2010). By definition, small populations of patients with rare conditions may make traditional recruitment targets infeasible and may require creativity in designing valid trials. Experts may design lower powered studies, increase the target significance level, consider alternative or surrogate outcomes, or even develop new methodology for small samples (Parmar et al. 2016,Billingham et al. 2016). Orphan drugs trials, while usually relatively small, have sometimes, but not always (Orfali et al. 2012), found to suffer from methodological shortcomings, such as lack of blinding or randomization (Kesselheim et al. 2011, Bell & Smith 2014). Furthermore, determination of the value of a drug is complex, based on factors such as disease prevalence, severity, mortality, morbidity, treatment benefit and safety, and cost effectiveness (Paulden et al. 2015).
3.1.3 Minimizing Overall Error

Mudge et al. (2012) recommend to consider the two types of error together and minimize the overall error rate. Two approaches include averaging the errors themselves and calculating the overall cost based on individual error costs. Similar approaches are applied in fields such as climate science, where the optimal \( \alpha \) level is calculated via simulation (Kemp 2016). (It is worth noting that in Kemp's study, \( \alpha = 0.0052 \), which was close to the recommendation by Benjamin et al. 2017) Mudge et al. (2017) even argue that these methods can simplify analyses with multiple hypothesis tests, which also require judgment as described in section 3.3. More generally, Grieve (2015) analyze how overall minimization approaches relate to the likelihood principle and Bayesian decision making. Implications of minimizing the overall error rate for research outcomes are modeled in Miller & Ulrich (2016).

3.2 Expert Judgment in Alternative Frameworks: Non-Inferiority Studies

The two hypotheses (the null and alternative) tested in standard statistical methods serve different functions, described in the appendix (section 8). Sprenger (2018) elaborates on the asymmetry between the two hypotheses and the inherent value judgments implicit in Frequentist methods. In non-inferiority trials, the roles of the two hypotheses are switched. (See the appendix for more details.) Non-inferiority studies require care in their design, analysis, and interpretation (Mauri & D’Agostino Sr 2017, Fleming et al. 2011, Fleming 2008). Examples include justification of an appropriate comparison group, choice of endpoint and non-inferiority margins “based on statistical reasoning and clinical judgment” (Fleming 2008), and careful a priori design to handle known challenges (Mauri & D’Agostino Sr 2017, D’Agostino Sr et al. 2003) Rehal et al. (2016) found that statistical recommendations are unclear and those that exist aren’t necessarily followed.
3.3 Multiple Comparisons

While testing multiple hypotheses in a single study is common, how and whether to adjust each test in light of the others is a non-trivial question. In certain settings, such as exploratory research, adjustment may be optional (Wason et al. 2014, Li et al. 2017, Gelman et al. 2012). Often, whether or not adjustment is needed may require judgment from a statistical expert about relationships between questions of interest (e.g. correlations or other measures of association). For example, tests comparing the effects of distinct treatments to a single control may not need adjustment, but tests comparing repeated measurements or multiple outcomes likely do (Candlish et al. 2017, Li et al. 2017, Wason et al. 2014).

Even when adjustment for multiple comparisons is generally agreed to be important, expertise is required in specifying an appropriate procedure for a particular problem. One issue to consider first is the number of tests for which to adjust (primary questions of interest only, primary and secondary questions, all planned tests, etc.). This issue requires input from both statisticians and subject matter experts about research priorities. The choice of adjustment method depends on the desired balance of type I and II errors, as discussed in Section 3.1. Additional considerations for multiple testing are provided elsewhere (Li et al. 2017, Wason et al. 2014, Alosh & Huque 2009, Proschan & Waclawiw 2000).

3.4 Should a binary decision be made at all?

The \( p \)-value was designed as an informal measure of the evidence that could cast doubt on the null hypothesis, not to inform a binary choice. Fisher (1955) objected to the decision theory framework of Neyman & Pearson (1928). Christensen (2005) describes a Fisherian view that, “an \( \alpha \) level should never be chosen; that a scientist should simply evaluate the evidence embodied in the \( p \)-value.” The modern use of \( p \)-values arguably has lost the emphasis on scientific judgment and extrapolates far beyond their intended use. Commentaries abound on the scientific and moral implications of the abuse of the \( p \)-value (Goodman 2016, Ziliak & McCloskey 2009, Steel et al. 2013, Pittenger 2001, Nickerson 2000, Folger 1989). This section discusses judgment of \( p \)-values outside of the decision-theory paradigm.

While some authors (e.g. Cumming 2014) and journals (see section 6.3) recommend
abandoning p-values entirely, others argue that alternatives aren’t necessarily better (Murtaugh 2014, Macnaughton n.d., Ionides et al. 2017, de Valpine 2014). Instead, p-values could be used in ways other than as sole decision-makers.

First, “borderline” values in either direction could be judged accordingly (Cohen 2011) with replication studies encouraged to reevaluate the findings. Given the nominal significance level $\alpha$ is an arbitrary choice, there is little qualitative difference between $p$-values in the interval $(\alpha - \epsilon, \alpha + \epsilon)$, where $\epsilon$ is a small number much smaller than the significance level. Yet, many researchers, even statisticians (McShane & Gal 2017) reject null hypotheses corresponding to $p$-values in the lower half of the interval and fail to reject null hypotheses corresponding to $p$-values in the upper half. While intuitively, there is little difference between $p$-values of 0.049 and 0.051, papers associated with the former are far more likely to be published than papers associated with the latter.

Additionally, $p$-values can be considered just one piece of evidence to be evaluated with other factors, such as study design quality and effect size (Spurlock 2017, Wasserstein & Lazar 2016). In fact, the United States Supreme Court issued an opinion to use statistical information as evidence rather than decision tools with blunt cut-offs (Liptak 2011).

4 Expert Judgment in Bayesian Methods

To mitigate problems with Frequentist inference, a large portion of the statistical community recommends greater engagement with Bayesian methods (Held & Ott 2018, Page & Satake 2017, Goldstein et al. 2006). The Bayesian paradigm enables more intuitive interpretations that directly answer questions of interest (Brownstein et al. 2018, Goldstein et al. 2006). The vast collection of literature on Bayesian methods extends far beyond the scope of the present paper. Appendix 8.2 provides a brief overview of the Bayesian paradigm.

It is well known that Bayesian statistics requires judgment in the choice of a prior distribution and that results may be sensitive to the choice of a prior (Gelman & Hennig 2017, Gelman et al. 2014). Therefore, careful consideration should be paid to the specification of the prior and its parameters. Determination of a prior requires input from both statisticians and experts in the fields of application. The focus of this section is an overview of prior development in Bayesian methods with emphasis on expert judgment and choice.
4.1 Expert Judgment in Choice of a Prior

Priors can be chosen in a variety of ways. Certain “standard” priors may be chosen for mathematical elegance, computational simplicity, or posterior robustness. Classes of priors, such as non-informative priors, conjugate families, and reference priors, are described extensively elsewhere (Bernardo 1979, Berger & Bernardo 1992, Syversveen 1998, Berger et al. 2009, DeGroot 2005, Fraser et al. 2010). More commonly, prior distributions and parameters may be determined from pilot data or other values from the literature. Details on the field of empirical Bayes, involving priors derived based on the current data, are found in chapter 5 of Carlin & Louis (2008). Little (2011) describes a hybrid between Bayesian and Frequentist methods called calibrated Bayes, in which Frequentist methods help determine the prior for a Bayesian model, which is used for inference. Priors for Bayesian models can also be chosen to coincide with Frequentist procedures (Daià & Ghosh 1995).

4.1.1 Expertise in Prior Elicitation

Alternatively, experts can directly develop and calibrate a prior distribution for their project. The process, called elicitation, is detailed in numerous places, such as Morgan (2014), O’Hagan et al. (2006), Garthwaite et al. (2005) and even in this special issue (O’Hagan 2018). In brief, a facilitator carefully works with content experts to quantify their judgment about parameters of interest and either directly uses or transforms the elicited densities into prior distributions for the model. Elicitation, which falls into the subjective Bayesian paradigm (Goldstein et al. 2006), is used in a variety of fields, such as clinical trials, environmental modeling economics and ecology (Mason et al. 2017, Krueger et al. 2012, O’Hagan 2012, Kuhnert et al. 2010, Martel et al. 2009).

Johnson et al. (2010) urge examination of elicitation methods for validity and reliability. Morgan (2014) details uses and abuses of elicitation, and Heitjan (2017) provides recent criticisms. While bias may be induced if elicitation is done poorly (Kynn 2008), elicitation can be used to estimate and mitigate bias (Turner et al. 2009). In fact, expert judgment and elicitation is critical to modern practice in environmental modeling (Krueger et al. 2012).
4.2 Model Checking in the Bayesian Paradigm

Regardless of the type of prior chosen for modeling, Gelman & Shalizi (2013) and Sprenger (2018) stress the importance of checking that the observed data matches the prior well enough for the application and considering alternative action if it diverges in key ways. For another example and discussion of implications for clinical trials when the prior doesn’t match the observed data, please see Brownstein et al. (2018).

5 Expert Judgment in Interpretation

Interpretation of results involves both knowledge of the statistical modeling process and broader experience with the subject matter. For example, to understand the results of a t-test of the null hypothesis that means in two populations are equal, one must first understand the basics behind hypothesis testing in general as well as assumptions and considerations specific to the two-sample t-test, such as normally distributed populations or large samples. Misinterpretation at this stage, such as thinking a $p$-value means something it doesn’t, or otherwise wrongly acting on a basic definition (confidence intervals, odds ratios, areas under ROC curves, etc.), are straightforward to correct with training in statistical literacy, as discussed in section 6.1. Of course, interpretation requires more than simply correctly applying definitions. Contextualizing results requires content expertise within the scientific field. Both types of expertise are required for other aspects, such as planning an analysis and evaluating potential bias, to name a few. This section focuses on examples including post-hoc difficulties with the planned analysis and missing data. Additional biases in interpretation of scientific evidence are examined in Kaptchuk (2003).

Choices abound in designing a study, from the experimental design to the statistical analysis and interpretation; these choices require scientific judgment. (Please see Brownstein et al. 2018 for details). Ideally, content experts and statisticians should jointly define and prioritize scientific questions and plan statistically valid, practical, and interpretable methods to answer them. Examples requiring judgment include decisions on the number and type of primary questions of interest, defining practically significant effect sizes on which to base sample size calculations, and whether and when to conduct interim
analyses.

A statistical analysis plan (SAP) is a formal document written prior to a study that describes the planned protocol. Adams-Huet & Ahn (2009) provide guidance for clinicians to work with statisticians in writing SAPs. Detailed SAPs are important to safeguard teams from changing their analyses after seeing the data (Finfer & Bellomo 2009). Such changes bias results and reduce reliability and validity.

Even with a well-written SAP, difficulties may arise. For instance, interpretation of the SAP for adjudication of complex potential events may be difficult. In one example, a single complex case in a clinical trial required an extensive investigation to interpret the statistical analysis plan properly (Gibson et al. 2017). In fact, the status of the single patient determined whether or not the $p$-value fell below the a priori defined significance level! When possible, SAPs can include strategies for dealing with missing or ambiguous data, such as sensitivity analyses, multiple imputation.

Another area which may necessitate careful statistical consideration is in the presence of excessive missing data. While a full review of missing data is outside of the scope of this paper, readers may consult Little & Rubin (2014). Briefly, analysis with missing data requires consideration of potential relationships between missing observations and the variables in the study. Standard methods can be complicated if missing data arises in non-random ways. As an illustrative example, the gold standard diagnosis of temporomandibular disorders (TMD) requires an invasive examination by an expert dentist (Dworkin 2010); missing data arises when subjects who are suspected to be new cases fail to complete the examination (Slade et al. 2013). The research team showed that the proportion of missing data was large, the assumptions of standard methods were violated, and new methods were required for modeling in the presence of the missing data (Brownstein et al. 2015, Bair et al. 2013). Recommendations, detailed further in section 6, include experimental protocols to check for and minimize missing data during data collection, and a priori plans (including existing methods or development of new methods) to handle analyses with missing data.
6 Recommendations from the Literature

Commentary on the role of statistical inference in the reproducibility crisis includes calls to action for members of the scientific community, including educational reform, improved standards for publication and funding and incentives for scientists. Section 6.1.1 details recommendations for statistics education, especially regarding teaching statistical inference to students not planning to become statisticians. Section 6.1.2 focuses on how to train statisticians with an eye toward improving reproducibility. Sections 6.2, 6.3, and 6.4 synthesize recommendations for stakeholders to facilitate better research practice for all fields.

6.1 Educational Interventions

The need for better statistical training in many fields has been discussed (Ogino & Nishihara 2016, Sørensen & Rothman 2010, Peng 2015). Other fields are recognizing the importance of detail at each step of the research progress (Shipworth & Huebner 2018). Indeed, Crane & Martin (2018) claim that statistical training can “[empower] scientists to make sound judgements.” Educational programs are discussed for scientists inside and out of statistics.

6.1.1 Educational Reform: Teaching Statistics for Practitioners

Despite the tendency for misuse among practitioners (Colquhoun 2017, Gardenier & Resnik 2002) and even statisticians (McShane & Gal 2017), basic statistical inference procedures and the accompanying $p$-values remain ubiquitous. One facilitator of the presence of significance tests is their central role in introductory courses and ease of implementation in statistical software (Searle 1980, Dallal 1990). In fact, the “cook-book” approach to statistical analysis commonly taught in introductory courses (Gigerenzer 2004) tempts researchers to apply basic procedures without investigating model diagnostics (Steel et al. 2013) or considering the appropriateness of the procedures for particular applications.

Crane & Martin (2018) and Brown & Kass (2009) stress that students in statistics courses should learn how to think critically. Furthermore, “statistical practice is complex, relying on nuanced principles honed through years of experience” (Crane & Martin...
In writing their guidelines to weed researchers, Onofri et al. (2010) make the following statement, “We would like to reinforce the idea that statistical methods are not a set of recipes whose mindless application is required by convention; each experiment or study may involve subtleties that these guidelines cannot cover.”

Paralleling the discussion for research, curriculum reform discussions are ongoing (Gould, Wild, Baglin, 2018; Park, 2018; Gould, Peng, Kreuter, Pruim, Witmer & Cobb, 2018). Example remedies for teaching students to use better judgment include increased focus on effect sizes and confidence intervals (Calin-Jageman, 2017; Fritz et al., 2012), greater attention to the Bayesian paradigm (Page & Satake, 2017), and practical demonstrations (Park, 2018; Mitchell, 2018).

Beyond the classroom, recommendations for biomedical scientists include programs on experimental design for both students and mentors, emphasis on methodology in journal clubs, developing continuing education materials as new methods arise, and increasing training for peer reviewers (Casadevall & Fang, 2016). It is especially important to better train peer reviewers in all fields to enforce statistical standards in published literature (Peng, 2015). Scientists can pursue additional formal statistical training while in graduate school or postdoctoral programs through funded mechanisms, such as NIH T32 training grants. However, the implementation of these specialized programs requires care to prevent trainees from simply learning a small amount of statistical tools and more confidently using them even when the tools are inappropriate for future problems (Gelfond et al., 2011). Bell et al. (2013) recommend both improved statistical training for content experts and continued collaboration with biostatisticians throughout the research process.

### 6.1.2 Improved Training for Quantitative Scientists

Statisticians serve as collaborators in a wide variety of fields, some of which strongly recommend or even require a statistician on the team (Obremskey & Archer, 2011; Bell et al., 2013). Consequently, proper training for statisticians and other quantitative science is paramount. Recognizing the widespread need for biostatistics in research and health practice as early as the 1960s, the National Institutes of Health (NIH) created biostatistics training programs (Hemphill, 1961). A more modern overview of training opportunities for biostatisticians is included in Kennedy et al. (2007).
Often working in fields outside of their primary educational training, statisticians should strive to continually learn about the subject-matter problems at hand (Brown & Kass 2009). Brown & Kass (2009) encourage graduate students in statistics to pursue a second program of study in joint or separate programs focused on the subject matter of their future research. Training opportunities, such as NIH K25, are available for this purpose, as described in Pickering et al. (2015). Formal training in a second field may not be feasible for all statisticians. However, applied practitioners could instead be encouraged to concentrate their collaborations in a small number of areas where they can slowly gain deeper familiarity with the scientific content, rather than to consult in a large number of disciplines where such depth is infeasible. In fact, this recommendation is often communicated in informal settings to junior faculty members with foci in applied statistics.

In addition, statisticians should hone their soft skills, especially listening, communication, and leadership skills (Gibson et al. 2017, Califf 2016). While listening and communication are obviously essential for collaboration, leadership training can provide statisticians with the tools and confidence to continually and actively shape the statistical validity of a project from its inception. In her March 2018 president’s corner article of AMSTAT news, LaVange (2018) reiterated the importance of leadership in biostatistics and announced the ASA’s vision for a new leadership initiative. As an example of the addition of soft skills to the curriculum, the biostatistics department at the University of North Carolina at Chapel Hill previously offered a course in leadership (LaVange et al. 2012).

6.1.3 Using Science to Develop Interventions for Scientists

A final idea about educational interventions is to study the research process scientifically and use the findings to improve training. Leek et al. (2017) summarized the state of affairs with a few apt quotes:

“The root problem is that we know very little about how people analyse and process information... We need to appreciate that data analysis is not purely computational and algorithmic — it is a human behaviour... We need more observational studies and randomized trials — more epidemiology on how people collect, manipulate, analyse, communicate and consume data. We can then use
To this end, in addition to actions described in section 6.2, funders could seek studies of scientific judgment with components to develop training materials based on the findings.

6.2 Funding Agencies

Due to the fact that funding is often necessary for research and highly valued or even required for promotions, funding agencies should take an active role in defining standards. To this end, the NIH defined guidelines for increased rigor, transparency and reproducibility (Hewitt et al. 2017, Collins & Tabak 2014). The NSF followed with guidelines to encourage data sharing and citation (Stan Ahalt et al. 2015). The ASA defined its own recommended actions for funding agencies to improve reproducibility, including funding mechanisms for training in reproducible research methods, development of reproducible software, replication studies (Broman et al. 2017). Actions by funding agencies should catalyze more widespread adoption of the recommendations for scientific rigor.

6.3 Journal Guidelines

Changes in publication standards percolate to the scientific practice of authors, as opined by Asher (1993) over two decades ago. In fact, this special *TAS* issue stems from high profile journal articles and rules. Namely, after Nuzzo (2014) bemoaned the ritual, often thoughtless, use of p-values, the editors of *Basic and Applied Social Psychology* banned p-values in manuscripts submitted thereafter (Trafimow & Marks 2015, Trafimow 2014). Recently, in *BMC Medical Research Methodology*, hypothesis testing has been “discouraged” (Hanin 2017). This section considers journal guidelines and their implications.

A recent set of journal guidelines for more careful publication practices (McNutt 2014) has been adopted by hundreds of journals thus far (Hewitt et al. 2017). Similarly, *BMC Medical Research Methodology* (Hanin 2017) advises authors that “Health care decisions [should be] based on... a combination of statistical and biomedical evidence.” Sørensen & Rothman (2010) propose that statistical training for journal editors could enable editors to better enforce statistical standards for publication in the journals that they oversee.
Journal guidelines include calls for transparency, including sharing of data and code, such as with pre-submission check-lists to ensure key aspects are considered (McNutt 2014). Other key aspects include reporting and justification of sample size calculations, randomization and blinding procedures, and inclusion and exclusion criteria (Asendorpf et al. 2013). Finally, emphasis can shift from inference to parameter estimation. (Asendorpf et al. 2013, McNutt 2014). These guidelines encourage authors to confront, acknowledge and justify the scientific judgments used in their studies and enable others to examine or reproduce their findings. The recommendations will likely improve rigor throughout science. Yet, implications from guidelines of some journals (Hanin 2017, Trafimow 2014) to avoid sharp thresholds for statistical inference is unclear, underlying a large portion of this special issue.

6.3.1 Transparency

Because statistics requires choices throughout the process, statistics has elements of subjectivity and objectivity, and Gelman & Hennig (2017) argue that the discussion of relative subjectivity and objectivity is distracting from solutions for best research practices, including transparency. Instead, paralleling the aim of journals to strive for transparency, Gelman & Hennig (2017) detail principles for which to strive during analysis, including, among others, acknowledgment and investigation of multiple perspectives, impartiality in decision making, and transparency in reporting. In brief, they argue that statistical judgment is inevitable, and therefore, the those judgments should be detailed and shared for future evaluation. Gibson (2017) stresses that statisticians should communicate “relative advantages” and disadvantages of each choice.

It has been shown that not only are descriptions of methods sections insufficiently detailed, but the quality of the reporting, methodology, and analysis are not necessarily associated with increased visibility (Nieminen et al. 2006). In addition, poor descriptions or unwillingness to share data may be associated with author concern over the robustness of results, especially for borderline results (Wicherts et al. 2011). This may be related to the need discussed in section 6.1 to increase statistical literacy broadly across other fields and professions. Statisticians can serve key roles as co-investigators by co-writing methods sections with sufficient detail for reproducibility, as peer reviewers who are specially trained...
to look for detail in reporting of study design, methods, and results, and as mentors and teachers who help others improve the completeness of the writing and peer review skills.

The ASA statement urges that “Proper inference requires full reporting and transparency.” Others similarly recommend increased transparency of study design and analysis (Gelman & Hennig 2017; Greenland et al. 2016; Peng 2015; McNutt 2014; Asendorpf et al. 2013), which would allow readers of the literature to evaluate findings in light of the strength of the methodology. There are even journals dedicated to reporting raw data (Data in Brief - Making Your Data Count 2014; Wang 2014), but data sharing is not yet widely adopted (Vasilevsky et al. 2017). Calls for transparency in research mirror calls nearly a decade ago for transparency in journalism (Weinberger 2009). However, others point out that too much a priori fixation and reporting of analysis plans stifles discovery (Poole 2010), and transparency alone cannot overcome poor data quality (Gelman 2017).

6.4 Interaction Between Groups

Clearly, changes to scientific and statistical standards will require coordination and buy-in from multiple groups concurrently, ensure that changes improve scientific practice as a whole (Ioannidis 2014; Sørensen & Rothman 2010). Asendorpf et al. (2013) include an excellent summary of recommended changes for these groups. Importantly, Smaldino & McElreath (2016) point out that incentives of various groups may conflict, such as funding agencies requesting publications from their grantees but not necessary checking the publication quality. As another example of potential misalignment of incentives, a seemingly simple solution advocated in Asendorpf et al. (2013) to increase sample sizes across science would improve reproducibility by improving power overall and likely subsequently increase the proportion of reported discoveries that are true. On the other hand, as the cost of each additional study participant may be large, increasing sample sizes may pose financial difficulties for already strained funding bodies. If total budgets were not increased, then requiring studies to be larger (and thus more expensive) could result in fewer projects funded. Such a consequence would negatively impact individual researchers who would face increased competition for comparatively fewer grants. Additional discussion of guidelines and their implications for various groups is found in Williams et al. (2018).
7 Conclusion

There is broad consensus that rigorous statistical practice and teaching require a mixture of technical knowledge, communication skills, and the ability to interpret data. The present paper highlights the less frequently discussed presence of choice and judgment in scientific and statistical practice. The universe of potential methods for analysis is large, and deciding among the options is challenging, even for experts (Gelman 2014). Properly synthesizing expert judgment along with quantitative tools is critical for high quality evidence-based research. Failure to appropriately capitalize on statistical and scientific judgment could further erode public trust in science (Spiegelhalter 2017, Saltelli & Funtowicz 2017) and policy (Sutherland et al. 2013, Weinberg & Elliott 2012). More importantly, scientists have an ethical obligation to conduct valid statistical analyses, which eventually have broader societal impacts (Shmueli 2017, Zook et al. 2017, Gelfond et al. 2011).

Rather than argue for a single solution to the problem of improving quantitative practice, this paper collects in a single document many of the thoughts on expert judgment and recommendations in statistics. (In a related piece submitted to this special issue, a strong statement on expert judgement in science with example applications by participants in the SSI is provided in Brownstein et al. 2018.) The present article gives the reader a starting point to understand the vast literature related to expert judgment in statistics. Viewpoints are highly diverse, indicating that there is much more work to be done toward developing concise, unified recommendations for improved methods. The author intends for the present paper to facilitate ongoing discussions of expert judgment and recommendations to improve statistical practice in the twenty-first century and beyond.

References

Adams-Huet, B. & Ahn, C. (2009), ‘Bridging clinical investigators and statisticians’, Journal of Investigative Medicine 57(8), 818–824.

Alosh, M. & Huque, M. F. (2009), ‘A flexible strategy for testing subgroups and overall population’, Statistics in Medicine 28(1), 3–23.
Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H. & Wicherts, J. M. (2013), ‘Recommendations for increasing replicability in psychology’, European Journal of Personality 27(2), 108–119.

Asher, W. (1993), ‘The role of statistics in research’, The Journal of Experimental Education 61(4), 388–393.

Bair, E., Brownstein, N. C., Ohrbach, R., Greenspan, J. D., Dubner, R., Fillingim, R. B., Maixner, W., Smith, S. B., Diatchenko, L., Gonzalez, Y. et al. (2013), ‘Study protocol, sample characteristics, and loss to follow-up: The oppera prospective cohort study’, The Journal of Pain 14(12), T2–T19.

Bell, M. L., Olivier, J. & King, M. T. (2013), ‘Scientific rigour in psycho-oncology trials: why and how to avoid common statistical errors’, Psycho-Oncology 22(3), 499–505.

Bell, S. A. & Smith, C. T. (2014), ‘A comparison of interventional clinical trials in rare versus non-rare diseases: an analysis of clinicaltrials.gov’, Orphanet journal of rare diseases 9(1), 170.

Benjamini, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C. et al. (2017), ‘Redefine statistical significance’, Nature Human Behaviour p. 1.

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 (Methodological) pp. 289–300.

Berger, J. O. & Bernardo, J. M. (1992), ‘On the development of the reference prior method’, Bayesian statistics 4, 35–60.

Berger, J. O., Bernardo, J. M., Sun, D. et al. (2009), ‘The formal definition of reference priors’, The Annals of Statistics 37(2), 905–938.
Berger, J. O. & Berry, D. A. (1988), ‘Statistical analysis and the illusion of objectivity’, *American Scientist* 76(2), 159–165.

Bernardo, J. M. (1979), ‘Reference posterior distributions for bayesian inference’, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 113–147.

Bertolaso, M. & Sterpetti, F. (2017), ‘Evidence amalgamation, plausibility, and cancer research’, *Synthese* pp. 1–39.

Billingham, L., Malottki, K. & Steven, N. (2016), ‘Research methods to change clinical practice for patients with rare cancers’, *The Lancet Oncology* 17(2), e70 – e80.

Braun, M. M., Farag-El-Massah, S., Xu, K. & Coté, T. R. (2010), ‘Emergence of orphan drugs in the united states: a quantitative assessment of the first 25 years’, *Nature Reviews Drug Discovery* 9(7), 519.

Broman, K., Cetinkaya-Rundel, M., Nussbaum, A., Paciorek, C., Peng, R., Turek, D. & Wickham, H. (2017), Recommendations to funding agencies for supporting reproducible research, in ‘American Statistical Association’, Vol. 2.

Brown, E. N. & Kass, R. E. (2009), ‘What is statistics?’, *The American Statistician* 63(2), 105–110.

Brownstein, N. C., Cai, J., Slade, G. D. & Bair, E. (2015), ‘Parameter estimation in Cox models with missing failure indicators and the OPPERA study’, *Statistics in Medicine* pp. 3984–3996.

Brownstein, N. C., Louis, T. A., O’Hagan, A. & Pendergast, J. (2018), ‘The role of expert judgement in statistical inference and evidence-based decision-making’, *The American Statistician* p. In press.

Byun, K. & Croucher, J. S. (2018), ‘Teaching statistics through the law’, *Teaching Statistics* pp. n/a–n/a. TST-Nov-17-1005.R2.

Califf, R. M. (2016), ‘Pragmatic clinical trials: Emerging challenges and new roles for statisticians’, *Clinical Trials* 13(5), 471–477.
Calin-Jageman, R. J. (2017), ‘After p values: The new statistics for undergraduate neuroscience education’, *Journal of Undergraduate Neuroscience Education* **16**(1), E1.

Candlish, J., Pate, A., Sperrin, M. & van Staa, T. (2017), ‘Evaluation of biases present in the cohort multiple randomised controlled trial design: a simulation study’, *BMC Medical Research Methodology* **17**(1), 17.

Carlin, B. P. & Louis, T. A. (2008), *Bayesian methods for data analysis*, CRC Press.

Carrier, M. (2010), ‘Scientific knowledge and scientific expertise: Epistemic and social conditions of their trustworthiness’, *Analyse & Kritik* **32**(2), 195–212.

Casadevall, A. & Fang, F. C. (2016), ‘Rigorous science: a how-to guide’, *mBio* **7**(6).

Christensen, R. (2005), ‘Testing fisher, neyman, pearson, and bayes’, *The American Statistician* **59**(2), 121–126.

Cohen, H. W. (2011), ‘P values: Use and misuse in medical literature’, *American Journal of Hypertension* **24**(1), 18–23.

Collins, F. S. & Tabak, L. A. (2014), ‘Nih plans to enhance reproducibility’, *Nature* **505**(7485), 612.

Colquhoun, D. (2017), ‘The reproducibility of research and the misinterpretation of p-values’, *Royal Society Open Science* **4**(12).

Crane, H. & Martin, R. (2018), ‘Is statistics meeting the needs of science?’.

Cumming, G. (2014), ‘The new statistics: Why and how’, *Psychological Science* **25**(1), 7–29. PMID: 24220629.

D’Agostino Sr, R. B., Massaro, J. M. & Sullivan, L. M. (2003), ‘Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics’, *Statistics in medicine* **22**(2), 169–186.

Daita, G. S. & Ghosh, J. K. (1995), ‘On priors providing frequentist validity for bayesian inference’, *Biometrika* **82**(1), 37–45.
Dallal, G. E. (1990), ‘Statistical computing packages: dare we abandon their teaching to others?’, *The American Statistician* 44(4), 265–266.

*Data in Brief - Making Your Data Count* (2014), *Data in Brief* 1, 5 – 6.

De Battisti, F., Ferrara, A. & Salini, S. (2015), ‘A decade of research in statistics: a topic model approach’, *Scientometrics* 103(2), 413–433.

de Valpine, P. (2014), ‘The common sense of p values’, *Ecology* 95(3), 617–621.

DeGroot, M. H. (2005), *Optimal Statistical Decisions*, John Wiley & Sons, Inc., chapter Conjugate Prior Distributions, pp. 155–189.

Dworkin, S. F. (2010), ‘Research diagnostic criteria for temporomandibular disorders: current status & future relevance’, *Journal of Oral Rehabilitation* 37(10), 734–743.

Ellis, P. D. (2010), *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*, Cambridge University Press.

Finfer, S. & Bellomo, R. (2009), ‘Why publish statistical analysis plans’, *Crit Care Resusc* 11(1), 5–6.

Fischer, A. & Ghelardi, G. (2016), ‘The precautionary principle, evidence-based medicine, and decision theory in public health evaluation’, *Frontiers in Public Health* 4, 107.

Fisher, R. (1955), ‘Statistical methods and scientific induction’, *Journal of the Royal Statistical Society. Series B (Methodological)* 17(1), 69–78.

URL: http://www.jstor.org/stable/2983785

Fjelland, R. (2016), ‘When laypeople are right and experts are wrong: Lessons from love canal’, *HYLE–International Journal for Philosophy of Chemistry* 22, 105–125.

Fleming, T. R. (2008), ‘Current issues in non-inferiority trials’, *Statistics in medicine* 27(3), 317–332.

Fleming, T. R., Odem-Davis, K., Rothmann, M. D. & Li Shen, Y. (2011), ‘Some essential considerations in the design and conduct of non-inferiority trials’, *Clinical Trials* 8(4), 432–439.
Folger, R. (1989), ‘Significance tests and the duplicity of binary decisions.’.

Francis, G. (2017), ‘Equivalent statistics and data interpretation’, Behavior Research Methods 49(4), 1524–1538.

Fraser, D., Reid, N., Marras, E. & Yi, G. (2010), ‘Default priors for bayesian and frequentist inference’, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(5), 631–654.

Fritz, C. O., Morris, P. E. & Richler, J. J. (2012), ‘Effect size estimates: current use, calculations, and interpretation.’, Journal of experimental psychology: General 141(1), 2.

Gardenier, J. & Resnik, D. (2002), ‘The misuse of statistics: Concepts, tools, and a research agenda’, Accountability in Research 9(2), 65–74. PMID: 12625352.

Garthwaite, P. H., Kadane, J. B. & O’Hagan, A. (2005), ‘Statistical methods for eliciting probability distributions’, Journal of the American Statistical Association 100(470), 680–701.

Gelfond, J. A., Heitman, E., Pollock, B. H. & Klugman, C. M. (2011), ‘Principles for the ethical analysis of clinical and translational research’, Statistics in medicine 30(23), 2785–2792.

Gelman, A. (2014), ‘How do we choose our default methods’, Past, Present, and Future of Statistical Science pp. 293–301.

Gelman, A. (2017), ‘Ethics and statistics: Honesty and transparency are not enough’, CHANCE 30(1), 37–39.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2014), Bayesian data analysis, Vol. 2, CRC press Boca Raton, FL.

Gelman, A. & Hennig, C. (2017), ‘Beyond subjective and objective in statistics’, Journal of the Royal Statistical Society: Series A (Statistics in Society) 180(4), 967–1033.

Gelman, A., Hill, J. & Yajima, M. (2012), ‘Why we (usually) don’t have to worry about multiple comparisons’, Journal of Research on Educational Effectiveness 5(2), 189–211.
Gelman, A. & Shalizi, C. R. (2013), ‘Philosophy and the practice of bayesian statistics’, *British Journal of Mathematical and Statistical Psychology* **66**(1), 8–38.

Gibson, C. M., Goldhaber, S. Z., Cohen, A. T., Nafee, T., Hernandez, A. F., Hull, R., Korjian, S., Daaboul, Y., Chi, G., Yee, M. et al. (2017), ‘When academic research organizations and clinical research organizations disagree: Processes to minimize discrepancies prior to unblinding of randomized trials’, *American heart journal* **189**, 1–8.

Gibson, E. W. (2017), ‘Leadership in statistics: Increasing our value and visibility’, *The American Statistician* **0**(ja), 00–00.

Gigerenzer, G. (2004), ‘Mindless statistics’, *The Journal of Socio-Economics* **33**(5), 587 – 606. Statistical Significance.

Goldstein, M. et al. (2006), ‘Subjective bayesian analysis: principles and practice’, *Bayesian analysis* **1**(3), 403–420.

Gómez, J. H., Marquina, V. & Gómez, R. (2013), ‘On the performance of usain bolt in the 100 m sprint’, *European journal of physics* **34**(5), 1227.

Goodman, S. N. (2016), ‘Aligning statistical and scientific reasoning’, *Science* **352**(6290), 1180–1181.

Gould, R., Peng, R. D., Kreuter, F., Pruim, R., Witmer, J. & Cobb, G. W. (2018), Challenge to the established curriculum: A collection of reflections, in ‘International Handbook of Research in Statistics Education’, Springer, pp. 415–432.

Gould, R., Wild, C. J., Baglin, J., McNamara, A., Ridgway, J. & McConway, K. (2018), *Revolutions in Teaching and Learning Statistics: A Collection of Reflections*, Springer International Publishing, Cham, pp. 457–472.

Greenland, S. & Poole, C. (2010-2011), ‘Problems in common interpretations of statistics in scientific articles, expert reports, and testimony’, *Jurimetrics* **51**, 113.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. & Altman, D. G. (2016), ‘Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations’, *European Journal of Epidemiology* **31**(4), 337–350.
Gribbin, J. & White, M. (2016), *Stephen Hawking: a life in science*, Pegasus Books.

Grieve, A. P. (2015), ‘How to test hypotheses if you must’, *Pharmaceutical Statistics* 14(2), 139–150.

Hanin, L. (2017), ‘Why statistical inference from clinical trials is likely to generate false and irreproducible results’, *BMC medical research methodology* 17(1), 127.

Heitjan, D. F. (2017), ‘Commentary on mason et al.’, *Clinical Trials* 14(4), 368–369.

Held, L. & Ott, M. (2018), ‘On p-values and bayes factors’, *Annual Review of Statistics and Its Application* 5(1), null.

Hemphill, E. C. (1961), ‘The probable impact of the nih training grant program on the future supply of biostatisticians’, *American Journal of Public Health and the Nations Health* 51(12), 1775–1779.

Hewitt, J. A., Brown, L. L., Murphy, S. J., Grieder, F. & Silberberg, S. D. (2017), ‘Accelerating biomedical discoveries through rigor and transparency’, *ILAR Journal* 58(1), 115–128.

Ioannidis, J. P. (2005), ‘Why most published research findings are false’, *PLoS medicine* 2(8), e124.

Ioannidis, J. P. (2014), ‘How to make more published research true’, *PLoS medicine* 11(10), e1001747.

Ionides, E. L., Giessing, A., Ritov, Y. & Page, S. E. (2017), ‘Response to the asa’s statement on p-values: Context, process, and purpose’, *The American Statistician* 71(1), 88–89.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T. & Feldman, B. M. (2010), ‘Methods to elicit beliefs for bayesian priors: a systematic review’, *Journal of clinical epidemiology* 63(4), 355–369.

Johnson, V. E. (2013), ‘Revised standards for statistical evidence’, *Proceedings of the National Academy of Sciences* 110(48), 19313–19317.
Johnson, V. E., Payne, R. D., Wang, T., Asher, A. & Mandal, S. (2017), ‘On the reproducibility of psychological science’, *Journal of the American Statistical Association* **112**(517), 1–10.

Kaptchuk, T. J. (2003), ‘Effect of interpretive bias on research evidence’, *Bmj* **326**(7404), 1453–1455.

Kemp, D. B. (2016), ‘Optimizing significance testing of astronomical forcing in cyclostratigraphy’, *Paleoceanography* **31**(12), 1516–1531.

Kennedy, R. E., Yeatts, S. D., Archer, K. J., Gennings, C. & Ramakrishnan, V. (2007), ‘Opportunities for biostatistics students: Training and fellowship grants from the national institutes of health’, *The American Statistician* **61**(2), 120–126.

Kesselheim, A., Myers, J. & Avorn, J. (2011), ‘Characteristics of clinical trials to support approval of orphan vs nonorphan drugs for cancer’, *JAMA* **305**(22), 2320–2326.

Krueger, T., Page, T., Hubacek, K., Smith, L. & Hiscock, K. (2012), ‘The role of expert opinion in environmental modelling’, *Environmental Modelling & Software* **36**, 4–18.

Kuhnert, P. M., Martin, T. G. & Griffiths, S. P. (2010), ‘A guide to eliciting and using expert knowledge in bayesian ecological models’, *Ecology Letters* **13**(7), 900–914.

Kynn, M. (2008), ‘The ‘heuristics and biases’ bias in expert elicitation’, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**(1), 239–264.

Lakens, D. (2013), ‘Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas’, *Frontiers in Psychology* **4**, 863.

Lakens, D., Adolphi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E. et al. (2018), ‘Justify your alpha’, *Nature Human Behaviour* **2**(3), 168.

LaVange, L. (2018), ‘Building a leadership institute (from the ground up)’, *Amstat News* **489**, 3–4.
LaVange, L., Sollecito, W., Steffen, D., Evarts, L. & Kosorok, M. (2012), ‘Preparing biostatisticians for leadership opportunities’, *Amstat News* **416**, 5–6.

Leek, J., McShane, B. B., Gelman, A., Colquhoun, D., Nuijten, M. B. & Goodman, S. N. (2017), ‘Five ways to fix statistics’, *Nature* **551**(7682), 557–559.

Li, G., Taljaard, M., Van den Heuvel, E. R., Levine, M. A., Cook, D. J., Wells, G. A., Devereaux, P. J. & Thabane, L. (2017), ‘An introduction to multiplicity issues in clinical trials: the what, why, when and how’, *International Journal of Epidemiology* **46**(2), 746–755.

Liptak, A. (2011), ‘Supreme court rules against zicam maker’, *New York Times* **22**.

Little, R. (2011), ‘Calibrated bayes, for statistics in general, and missing data in particular’, *Statist. Sci.* **26**(2), 162–174.

Little, R. J. & Rubin, D. B. (2014), *Statistical analysis with missing data*, Vol. 333, John Wiley & Sons.

Macnaughton, D. B. (n.d.), ‘The p-value is best to detect effects’. [Online; accessed 13-July-2018].

URL: [https://matstat.com/macnaughton2017b.pdf](https://matstat.com/macnaughton2017b.pdf)

Marino, M. J. (2017), ‘How often should we expect to be wrong? statistical power, p-values, and the expected prevalence of false discoveries’, *Biochemical Pharmacology*.

Martel, M., Hernández, M. Á. N. & Polo, F. J. V. (2009), ‘Eliciting expert opinion for cost-effectiveness analysis: a flexible family of prior distributions’, *SORT-Statistics and Operations Research Transactions* **33**(2), 193–212.

Mason, A. J., Gomes, M., Grieve, R., Ulug, P., Powell, J. T. & Carpenter, J. (2017), ‘Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: application to the improve trial’, *Clinical Trials* **14**(4), 357–367.
Matthews, R., Wasserstein, R. & Spiegelhalter, D. (2017), ‘The asa’s p-value statement, one year on’, *Significance* **14**(2), 38–41.

Mauri, L. & D’Agostino Sr, R. B. (2017), ‘Challenges in the design and interpretation of noninferiority trials’, *New England Journal of Medicine* **377**(14), 1357–1367.

McNutt, M. (2014), ‘Journals unite for reproducibility’, *Science* **346**(6210), 679–679.

McShane, B. B. & Gal, D. (2017), ‘Statistical significance and the dichotomization of evidence’, *Journal of the American Statistical Association* **112**(519), 885–895.

Merriam-Webster Online Dictionary (2018a), ‘Statistician’. [Online; accessed 11-July-2018].
URL: https://www.merriam-webster.com/dictionary/statistician

Merriam-Webster Online Dictionary (2018b), ‘Statistics’. [Online; accessed 11-July-2018].
URL: https://www.merriam-webster.com/dictionary/statistics

Miller, J. & Ulrich, R. (2016), ‘Optimizing research payoff’, *Perspectives on Psychological Science* **11**(5), 664–691. PMID: 27694463.

Mitchell, P. (2018), ‘Teaching statistical appreciation in quantitative methods’, *MSOR Connections* **16**(2), 37–42.

Morey, R. D., Romeijn, J.-W. & Rouder, J. N. (2016), ‘The philosophy of bayes factors and the quantification of statistical evidence’, *Journal of Mathematical Psychology* **72**, 6–18.

Morgan, M. G. (2014), ‘Use (and abuse) of expert elicitation in support of decision making for public policy’, *Proceedings of the National Academy of Sciences* **111**(20), 7176–7184.

Motulsky, H. (2014), *Intuitive biostatistics: a nonmathematical guide to statistical thinking*, Oxford University Press, USA.

Mudge, J. F., Baker, L. F., Edge, C. B. & Houlanah, J. E. (2012), ‘Setting an optimal α that minimizes errors in null hypothesis significance tests’, *PLoS ONE* **7**(2), 1–7.

Mudge, J. F., Martyniuk, C. J. & Houlanah, J. E. (2017), ‘Optimal alpha reduces error rates in gene expression studies: a meta-analysis approach’, *BMC Bioinformatics* **18**(1), 312.
Murphy, K. R., Myors, B. & Wolach, A. (2014), *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*, Routledge.

Murtaugh, P. A. (2014), ‘In defense of p values’, *Ecology* 95(3), 611–617.

Neyman, J. & Pearson, E. S. (1928), ‘On the use and interpretation of certain test criteria for purposes of statistical inference: Part i’, *Biometrika* pp. 175–240.

Nickerson, R. S. (2000), ‘Null hypothesis significance testing: a review of an old and continuing controversy.’, *Psychological methods* 5(2), 241.

Nieminen, P., Carpenter, J., Rucker, G. & Schumacher, M. (2006), ‘The relationship between quality of research and citation frequency’, *BMC Medical Research Methodology* 6(1), 42.

Nuzzo, R. (2014), ‘Scientific method: statistical errors’, *Nature News* 506(7487), 150.

Obremskey, W. T. & Archer, K. R. (2011), ‘Getting started in an academic setting’, *Journal of orthopaedic trauma* 25, S124–S127.

Ogino, S. & Nishihara, R. (2016), ‘All biomedical and health science researchers, including laboratory physicians and scientists, need adequate education and training in study design and statistics’, *Clinical Chemistry* 62(7), 1039–1040.

O’Hagan, A. (2012), ‘Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux’, *Environmental Modelling & Software* 36, 35–48.

O’Hagan, A. (2018), ‘Expert knowledge elicitation: Subjective, but scientific’, *The American Statistician* p. In press.

O’Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. & Rakow, T. (2006), *Uncertain judgements: eliciting experts’ probabilities*, John Wiley & Sons.

Onofri, A., Carbonell, E. A., Piepho, H.-P., Mortimer, A. M. & Cousens, R. D. (2010), ‘Current statistical issues in weed research’, *Weed Research* 50(1), 5–24.
Orfali, M., Feldman, L., Bhattacharjee, V., Harkins, P., Kadam, S., Lo, C., Ravi, M.,
Shringarpure, D., Mardekin, J., Cassino, C. et al. (2012), ‘Raising orphans: how clinical
development programs of drugs for rare and common diseases are different’, *Clinical
pharmacology & Therapeutics* **92**(2), 262–264.

Page, R. & Satake, E. (2017), ‘Beyond p values and hypothesis testing: Using the mini-
mum bayes factor to teach statistical inference in undergraduate introductory statistics
courses’, *Journal of Education and Learning* **6**(4), 254.

Park, R. (2018), ‘Practical teaching strategies for hypothesis testing’, *The American Statis-
tician* **0**(ja), 0–0.

Parmar, M. K. B., Sydes, M. R. & Morris, T. P. (2016), ‘How do you design randomised
trials for smaller populations? a framework’, *BMC Medicine* **14**(1), 183.

Paulden, M., Stafinski, T., Menon, D. & McCabe, C. (2015), ‘Value-based reimbursement
decisions for orphan drugs: a scoping review and decision framework’, *Pharmacoecono-
nomics* **33**(3), 255–269.

Peng, R. (2015), ‘The reproducibility crisis in science: A statistical counterattack’, *Signif-
icance* **12**(3), 30–32.

Pickering, C. R., Bast, R. C. & Keyomarsi, K. (2015), ‘How will we recruit, train, and retain
physicians and scientists to conduct translational cancer research?’, *Cancer* **121**(6), 806–
816.

Pittenger, D. J. (2001), ‘Hypothesis testing as a moral choice’, *Ethics & Behavior*
**11**(2), 151–162.

Poole, C. (2010), ‘A vision of accessible epidemiology’, *Epidemiology* **21**(5), 616–618.

Proshan, M. A. & Waclawiw, M. A. (2000), ‘Practical guidelines for multiplicity adjust-
ment in clinical trials’, *Contemporary Clinical Trials* **21**(6), 527–539.

Rehal, S., Morris, T. P., Fielding, K., Carpenter, J. R. & Phillips, P. P. J. (2016), ‘Non-
inferiority trials: are they inferior? a systematic review of reporting in major medical
Saltelli, A. & Funtowicz, S. (2017), ‘What is science’s crisis really about?’, Futures 91, 5 – 11. Post-Normal science in practice.

Schell, M. J. (2010), ‘Identifying key statistical papers from 1985 to 2002 using citation data for applied biostatisticians’, The American Statistician 64(4), 310–317.

Searle, S. R. (1989), ‘Statistical computing packages: Some words of caution’, The American Statistician 43(4), 189–190.

Shipworth, D. & Huebner, G. M. (2018), Designing research, in ‘Exploring Occupant Behavior in Buildings’, Springer, pp. 39–76.

Shmueli, G. (2017), ‘Research dilemmas with behavioral big data’, Big data 5(2), 98–119.

Slade, G. D., Fillingim, R. B., Sanders, A. E., Bair, E., Greenspan, J. D., Ohrbach, R., Dubner, R., Diatchenko, L., Smith, S. B., Knott, C. et al. (2013), ‘Summary of findings from the oppera prospective cohort study of incidence of first-onset temporomandibular disorder: implications and future directions’, The Journal of Pain 14(12), T116–T124.

Smaldino, P. E. & McElreath, R. (2016), ‘The natural selection of bad science.’, Royal Society Open Science 3(9), 160384.

Sørensen, H. T. & Rothman, K. J. (2010), ‘The prognosis for research’.

Spiegelhalter, D. (2017), ‘Trust in numbers’, Journal of the Royal Statistical Society: Series A (Statistics in Society) 180(4), 948–965.

Sprenger, J. (2018), ‘The objectivity of subjective bayesianism’, European Journal for Philosophy of Science pp. 1–20.

Spurlock, D. (2017), ‘Beyond p<. 05: Toward a nightingalean perspective on statistical significance for nursing education researchers’, Journal of Nursing Education 56(8), 453–455.
Stan Ahalt, R., Couch, A., Ibanez, L. & Ray Idaszak, R. (2015), ‘Nsf workshop on supporting scientific discovery through norms and practices for software and data citation and attribution’.

Steel, E. A., Kennedy, M. C., Cunningham, P. G. & Stanovick, J. S. (2013), ‘Applied statistics in ecology: common pitfalls and simple solutions’, *Ecosphere* 4(9), 1–13.

Sutherland, W. J., Burgman, M. A. et al. (2015), ‘Use experts wisely’, *Nature* 526(7573), 317–318.

Syversveen, A. R. (1998), ‘Non-informative bayesian priors: Interpretation and problems with construction and applications’.

**URL:** https://www.ime.unicamp.br/veronica/MI402/Randi21998.pdf

Tibshirani, R. (1989), ‘Noninformative priors for one parameter of many’, *Biometrika* 76(3), 604–608.

Trafimow, D. (2014), ‘Editorial’, *Basic and Applied Social Psychology* 36(1), 1–2.

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

Turner, R. M., Spiegelhalter, D. J., Smith, G. & Thompson, S. G. (2009), ‘Bias modelling in evidence synthesis’, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172(1), 21–47.

Vasilevsky, N. A., Minnier, J., Haendel, M. A. & Champieux, R. E. (2017), ‘Reproducible and reusable research: are journal data sharing policies meeting the mark?’.

Wang, H.-R. (2014), “‘publish or perish”: Should this still be true for your data?’, *Data in Brief* 1, 85 – 86.

Wason, J. M. S., Stecher, L. & Mander, A. P. (2014), ‘Correcting for multiple-testing in multi-arm trials: is it necessary and is it done?’, *Trials* 15(1), 364.

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.
8 Appendix: The Basics and Philosophy of Hypothesis Testing

Hypothesis testing compares two hypotheses, called the null and the alternative. Two paradigms are presented for comparing the hypotheses using data collected in scientific experiments. Other overviews are found elsewhere, e.g. Greenland & Poole (2010-2011).
8.1 The Frequentist Paradigm

In Frequentist hypothesis testing, the null hypothesis is considered the default, assumed to be true unless the data casts doubt on it. The alternative hypothesis is often what a content expert may hope to conclude is true by casting doubt on the null hypothesis. For example, the developer of a new cancer treatment may compare changes in tumor size for patients randomized to either their new treatment or a placebo. In this case, the null hypothesis is that average change in tumor size is identical for patients randomized to both interventions, and the alternative hypothesis is that the change in tumor size differs based on whether the patients were randomized to the new drug or a placebo. Put another way, the drug developer hopes to convince regulatory bodies that their drug is effective by casting doubt on the default theory that the effect of the drug is identical to the effect of a placebo.

If one assumes that the two (mutually exclusive) hypotheses exhaust reasonable possibilities, then exactly one of the hypotheses should be true and the other should be false. The practitioner either rejects or fails to reject the null hypothesis based on the evidence in the experiment. While the hope is that the conclusion aligns with the (unknown) truth, two errors can take place. The first, called a type I error, occurs when the null hypothesis is true, but the practitioner rejects the null hypothesis. The second, called a type I error, occurs when the alternative hypothesis is true, but the practitioner fails to reject the null hypothesis. Type I errors can be conceptualized as false positives, while type II errors can be considered false negatives. The significance level is the probability of a type I error. Power, equal to the probability of the complement of a type II error, is the probability of (correctly) rejecting the null hypothesis when the alternative hypothesis is true.

For a fixed sample size, type I and type II errors are inversely related (when one increases, the other decreases), and the appropriate balance should be decided carefully. Typically, type I error is set at an appropriate value to minimize the occurrence of false discoveries, which are typically considered worse than missing a true discovery. The sample size for the experiment is then set based on both the significance level (e.g. 5%) and a desired level of power (e.g. 80%) under a specified realization of the alternative hypothesis. Alternatively, often in the presence of extremely limited resources, power is calculated based on the fixed significance level and the maximum sample size considered feasible.
The significance level is arbitrary, but 5% has been the most frequently adopted value by tradition (Ziliak & McCloskey 2009). The idea is that the probability of falsely rejecting the null hypothesis, e.g. concluding that an association of interest in the study is present in the population when in reality no such association exists, needs to be minimized. Otherwise, not only will the original findings mislead the research community, but future research will be designed based on the previous (incorrect) findings.

As detailed in Motulsky (2014), one can conceive of the Frequentist hypothesis framework as similar to a jury evaluating the evidence in a criminal trial: the defendant is presumed by default not to be guilty of the crime and is only deemed to be guilty of the crime if there is sufficient evidence of guilt beyond a reasonable doubt. In the courtroom, wrongful convictions are considered worse than failing to convict a person who committed a crime in the presence of insufficient evidence. Similarly, type I errors are typically set at lower rates (e.g. 5%) than type II errors (10% or 20%). Additional connections between statistics and law and examples for teaching are provided in Byun & Croucher (2018).

It is important to note that the null and alternative hypotheses serve different functions and are not interchangeable. A common misconception is to interpret a failure to reject the null hypothesis as “accepting the null hypothesis.” Critically, p-values do not quantify the evidence in favor of the null hypothesis, analogous to the fact that criminal trials never claim to prove the innocence of a client and merely deliver verdicts without convictions as as “not guilty” (Motulsky 2014). In some cases, the order of the hypotheses may need to be flipped, such as for non-inferiority trials, where the goal is to disprove the notion that a generic compound differs from the original version in favor of the alternative that the two compounds are sufficiently similar. Thus the framework, analysis, and interpretation of non-inferiority studies differ markedly from the standard hypothesis testing framework.

The body of the paper discusses an inverse relationship between type I and type II errors when the sample size is fixed. Exacerbating the trade-off is the fact that most research aims to answer more than one question of interest, and therefore, a significant portion of the literature includes more than one hypothesis test. The family-wise error rate, defined as the probability of having at least one type I error among a fixed number of tests, is known to be higher than the nominal significance level of any single test, as shown in
Li et al. (2017). In fields, such as genomics, with a large number of planned hypothesis tests, adjusted methods aim to control the false discovery rate, defined as the proportion of the null hypotheses that are true among all tests for which the null hypotheses is rejected (Benjamini & Hochberg 1995). It is important to note that family-wise error rates and false discovery rates are not equivalent. Calculation of the false discovery rate utilizes Bayes’ rule and necessitates a judgment on the proportion of null hypotheses that are true.

8.2 The Bayesian Paradigm

As described in Section 8.1, the Frequentist paradigm answers questions about the data assuming that one of the competing hypotheses is true. In particular, a $p$-value quantifies the evidence an experiment produces assuming that the null hypothesis is true. In reality, even after the experiment is complete, the research team cannot definitively answer the question of whether or not the null hypothesis is true. It is intuitively more desirable to quantify and compare the posterior probabilities of each hypothesis conditional on the data from the experiment. The Bayesian paradigm provides methodology for this framework.

Consider the example from Section 8.1 of a new oncology drug. Under the Bayesian paradigm, a quantity of interest is the posterior probability of the null hypothesis that the drug and placebo change tumor size by amounts that are nearly identical (up to a range) based on the data from the patients in the trial and the previously hypothesized efficacy of the drug. Similarly of interest is the posterior probability that the drug reduces the tumor size by a greater margin than the placebo conditional on the observed data. The drug developer hopes to convince regulatory bodies that their drug is effective by directly calculating the posterior probability, based on the results of the trial, that there is a meaningful underlying difference in treatment response between their drug and a placebo.

Another key difference between Bayesian and Frequentist methodology is dependence of Bayesian posterior probabilities on a prior distribution. The prior specifies the probability that the null hypothesis is true (or false). Another quantity of interest in this paradigm is the Bayes factor, which is the ratio of the likelihood of the data under one hypothesis to the likelihood of the data under the other hypothesis. The philosophy behind the Bayes factor is detailed in Morey et al. (2016).
More generally, Bayesian methods involve specification for the parameters of interest. In the presence of nuisance parameters, i.e. parameters that are not of interest themselves but must be defined to answer the question of interest, this endeavor may be especially challenging (Fraser et al. 2010, Tibshirani 1989). Further details on the Bayesian paradigm and how to implement Bayesian models may be found in a variety of sources, such as textbooks by Gelman et al. (2014) and Carlin & Louis (2008).