The Future of Quantitative Inquiry in Education: Challenges and Opportunities

David Kaplan
Department of Educational Psychology
University of Wisconsin – Madison

In 2015, after a one-year grace period, the journal Basic and Applied Social Psychology (BASP) issued an editorial policy banning null hypothesis significance testing (NHST). The policy now allows the author(s) to submit papers that present results using NHST procedures, but prior to publication “…authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about significant differences or lack thereof, and so on).”(Trafimow and Marks, 2015, pg. 1).

The editorial is set up in a question and answer format. In response to the question of whether other inferential procedures such as Bayesian methods would be acceptable, Trafimow and Marks write “… with respect to Bayesian procedures, we reserve the right to make case-by-case judgments, and thus Bayesian procedures are neither required nor banned from BASP.” Although it is not clear what criteria the editors of BASP would use to judge the acceptability of a Bayesian analysis, the Bayesian paradigm does offer a clear direction away from adherence to NHST and this will be further articulated below.

Criticisms about NHST have existed for almost as long as the paradigm itself. Early concerns about NHST were expressed by Jeffreys (1961). More recent criticisms can be found in e.g. Cohen (1994), Gigerenzer et al. (2004), Wagenmakers (2007) and the volume by Harlow, Mulaik, and Steiger (1997), among many others. Nevertheless, the announcement of the ban by BASP was met with considerable discussion over various quantitative methods blogs as well as editorials from quantitative methodology journals.

What impact might a ban on NHST have for the future of quantitative inquiry in education? This paper examines the motivations for the ban by briefly describing the theory of NHST. Next, we focus on the precise interpretation of the p-value. We then discuss the challenges associated with NHST, and argue that the Bayesian inferential paradigm represents a constructive way forward for quantitative inquiry in education.

1 I am grateful to Daniel Bolt and Peter Steiner for valuable discussion on this topic. The opinions expressed are mine alone.
A critically important component of quantitative inquiry in education is inference and model building. Whether interest centers on examining school-based interventions via clustered randomized designs or complex predictive models using data from international large-scale educational assessments such as PISA or TIMSS, the goal, arguably is to quantify the evidence for our research hypotheses. To this end, a considerable amount of time is spent in introductory statistics courses laying the foundations of hypothesis testing, starting with Fisher (1971/1935) and culminating in the Neyman and Pearson (1928) framework.

An interesting aspect of NHST is that students (as well as many seasoned researchers) appear to have a very difficult time grasping its principles. Gigerenzer, Krauss, and Vitouch (2004) argued that much of the difficulty in grasping conventional hypothesis testing lies in the conflation of the approaches advocated by Fisher (1971/1935) and by Neyman and Pearson (1928).

Briefly, Fisher's approach to hypothesis testing requires specifying only the null hypothesis. For Fisher, the term “null” meant, the hypothesis to be “nullified”. A conventional significance level is chosen (usually the 5% level). Once the test is conducted, the result is either significant ($p < 0.05$) or it is not ($p \geq 0.05$). If the resulting test is significant, then the null hypothesis is rejected. However, if the resulting test is not significant, then no conclusion can be drawn.

In contrast to Fisher's ideas, the approach advocated by Neyman and Pearson requires that two hypotheses be specified -- the null and alternative hypothesis. By specifying two hypotheses, one can compute a desired tradeoff between two types of errors: Type I errors (the probability of rejecting the null when it is true), and Type II errors (the probability of not rejecting the null when it is false). The Neyman and Pearson approach is, in some sense, a decision-theoretic framework, providing information leading to an action taken by the researcher. Under the Neyman and Pearson approach, studies are ideally designed prior to data collection so as to minimize Type I or Type II errors depending on the goals of the research. The important point here is that the Neyman and Pearson approach is not a framework for quantifying evidence, it is a model for minimizing errors.

The conflation of Fisher and Neyman and Pearson hypothesis testing lies in the use and interpretation of the $p$-value. In Fisher's paradigm, the $p$-value is a matter of convention with the resulting outcome being based on the data. In the Neyman and Pearson paradigm the Type I and Type II error probabilities are determined prior to the experiment being conducted and refer to a consideration of the cost of making one or the other error. However, even a casual perusal of the top journals in education and the social sciences generally will reveal that this balance is virtually always ignored and a Type I error probability of 0.05 is used. The conventional 0.05 level itself being the result Fisher's

---

2 The term “experiment” is used here to refer to any systematic data collection exercise and encompasses observational studies, quasi-experiments, as well as randomized experiments.
experience with small agricultural experiments. To quote Gigerenzer, Krauss, and Vitouch (2004) “For Fisher, the exact level of significance is a property of the data, that is, a relation between a body of data and a theory; for Neyman and Pearson, $\alpha$ is a property of the test, not of the data. Level of significance and $\alpha$ are not the same thing.”

The problem with NHST, therefore, seems to lie with the interpretation of the $p$-value. To be precise, the $p$-value rests on a form of argumentation referred to as *reductio ad absurdum*. This form of argumentation attempts to establish an assertion by deriving an absurdity from its denial, thus arguing that the assertion must be accepted because its rejection would be indefensible. The $p$-value is thus properly interpreted as the probability of observing data at least as extreme as the data that was actually observed, computed under the assumption that the null hypothesis is true.

Let’s look at the interpretation of the $p$-value carefully. First, the $p$-value is based on data that were never observed. Specifically, the $p$-value is obtained by referencing the value of the observed test statistic (e.g. the $t$-test) based on the study at hand to hypothetically replicated data generated under the assumption that the null hypothesis is true. Second, the $p$-value is computed under the assumption that the null hypothesis is true. In most applications, the null hypothesis is taken to reflect no difference (or no effect, no relationship, etc.). To begin with, this hypothesis is never true in reality, and moreover, it is typically not the research question of interest. Thus, after Cohen (1994), researchers are typically testing a “nil” hypothesis that is hopefully rejected. However, it is important to emphasize that there is nothing within the NHST paradigm that requires testing the “nil” hypothesis. In fact, any theoretically justifiable hypothesis can serve as the null hypothesis.

So, to summarize, the $p$-value does not quantify evidence for a hypothesis, rather, it provides a measure of the probability of an outcome not actually observed, computed under the assumption of a null hypothesis (in fact, a “nil” hypothesis) that will not be true in any population. To quote Jeffreys (1961, pg. 385) “This seems a remarkable procedure”.

Obviously, misunderstandings of NHST and the associated $p$-value are not sufficient to ban its use. However, these misunderstandings combined with the clear and well-documented bias toward publishing only statistically significant results, has pushed authors to questionable practices, such as describing non-significant findings as “trending toward significance” -- a phrase, among many in use today, that has no basis in statistical theory. However, more seriously perhaps for the advancement of quantitative inquiry in education, the conventional the $p$-value does not quantify evidence for a hypothesis of interest, and this, arguably is what education researchers are after.

What then are the opportunities for hypothesis testing in education? First, it should be noted that NHST could be valuable when rigorous error control is desired. However, for practical applications of model building and evaluation, we argue that the Bayesian paradigm of statistical inference represents an internally consistent and coherent alternative to NHST -- an alternative that is now widely available to education
Researchers due to the development of computational algorithms applicable to Bayesian analysis. Classic texts on Bayesian statistics can be found in e.g. de Finetti (1974) and Savage (1954). A discussion of Bayesian statistics with applications to education research can be found in Kaplan (2014). An excellent treatment of Bayesian epistemology can be found in Howson and Urbach (2006).

Briefly, Bayesian inference treats probability as the language for encoding uncertainty about those elements of an analysis that are unknown; in particular, model parameters such as treatment effects or regression coefficients. Second, unknown parameters are assumed to be random variables described by probability distributions representing what is reasonable to believe about the parameters of interest. These probability distributions are referred to as prior distributions that can be elicited from personal subjective belief, expert opinion, and/or prior research. Third, through Bayes' theorem, prior distributions on the model parameters are combined with the probability model of interest (e.g. a regression model) to yield updated knowledge about the unknown parameters summarized in the so-called posterior distributions. Finally, the focus of model evaluation in Bayesian inference is based largely on predictive quality and not on goodness-of-fit, per se. Statistical significance testing, and all its “vestiges” (including NHST p-values) play virtually no role in Bayesian inference. Thus, Bayesian inference is a framework for learning from data. To quote Jerome Cornfield (cited in Savage, 1954), “[I]t is clear that it is not possible to think about learning from experience and acting on it without coming to terms with Bayes’ theorem.”

To be clear, Bayesian inference is not without its own internal set of controversies – the most salient of which is the choice of priors. Recall that priors can be obtained from personal subjective belief, expert opinion, and/or prior research. Eliciting priors, especially those based on subjective belief, is fraught with difficulties. Indeed, without careful techniques of elicitation and formal comparison among models with different priors, a researcher can skew his/her results toward their prior beliefs. The problem of elicitation is discussed in O’Hagan, et al. (2006). In response, many in the Bayesian world have sought so-called “objective priors” that retain the benefits of Bayesian framework for uncertainty quantification while at the same time letting the data “speak” as much as possible. This controversy has divided the Bayesian world into “subjectivists” (e.g. de Finetti, 1974) and “objectivists” (e.g. Berger, 2006). Rich methodological research is continuing to flow from these two schools within the Bayesian paradigm and it is important that those engaged in quantitative education inquiry become familiar with the debate (see e.g. Kaplan, 2014, Chapter 10, for a discussion of the debates within the Bayesian paradigm).

To summarize, Bayesian statistical inference offers a constructive alternative to NHST for quantitative inquiry in education, guiding education research toward evolutionary knowledge development and away from blind adherence to NHST. However, we do not believe that NHST should be banned. In fact, as noted above, there may be situations in

---

3 Bayesian inference can also treat the uncertainty in model building and selection through the method of Bayesian model averaging. This is beyond the scope of the paper.
which the kind of error control that is achieved by the proper use of NHST is desired. Nevertheless, we would argue that such situations are relatively rare in education research. Rather, often the goal of quantitative inquiry in education is to assess whether a hypothesis of interest is supported by the data in hand. Addressing this goal forms the basis of Bayesian inference; recognizing and explicitly accounting for all manner of uncertainty that can enter into an inquiry.

For a Bayesian approach to quantitative inquiry in education to move forward, several steps are at necessary: (a) Bayesian epistemology should be openly taught alongside NHST in introductory statistics classes, clearly defining for students the challenges and opportunities of each approach, (b) require that authors clearly warrant the choice of either NHST or Bayesian inference for their investigations, and (c) develop and demonstrate "best practices" in Bayesian inference and advocate for the adoption of these best practices by research journals and funding agencies. These steps would go a long way in improving quantitative inquiry in education.

References


