HYPOTHESIS TESTING IN PSYCHOLOGY: THROWING THE BABY OUT WITH THE BATHWATER? ®

Michael Granaas
University of South Dakota
USA

For many years null hypothesis testing (NHT) has been the dominant form of statistical analysis in psychology. It has also been subject to periodic criticisms from within the field of psychology. In the past decade these occasional criticisms have turned into a more or less steady stream which have lead some to call for an outright ban on NHT in psychology, while others have called for greater use of alternative procedures. The solution lies neither in banning NHT nor in relying solely on alternative procedures, but in "reforming" NHT, replacing a-theoretical null hypotheses with theoretically meaningful hypotheses. Such reform requires that training of researchers emphasize parameter estimation and the testing of theoretical models, an approach that exists in some areas of psychology and appears to be common in other sciences. Such an emphasis will help ensure that the statistical hypothesis being tested matches the substantive hypothesis of interest. I will discuss the changes that are occurring in psychology and propose further changes that are still needed.

INTRODUCTION

There have long been psychologists who are critical of the use of Null Hypothesis Testing (NHT) in psychology (e.g., Boring, 1919; Rozeboom, 1960). Despite periodic criticism of NHT during the 1960's, 70's, and 80's there was very little change in statistical practice (Cohen, 1994). Renewed criticism of NHT by psychologists during the 1990's (e.g., Cohen, 1994; Schmidt, 1996) lead to the publication of a special issue of *Psychological Science* (vol 8 (1)), and an edited volume (Harlow, Mulaik, & Steiger, Eds, 1997) presenting many of the arguments both pro and con regarding the practice and future of NHT in psychology.

Fueled by the high level of concern the American Psychological Association (APA) convened a task force to examine and make recommendations regarding statistical practice within the field of Psychology (Wilkinson & APA Task Force, 1999). While the final report of the Task Force appears to have reduced the volume of criticism, the controversy is far from over.

Nickerson (2000) has summarized and discussed the various reasons behind the criticism and concern surrounding the use of NHT in psychological research. Much of the debate focuses on the problems of misunderstanding and misinterpretation of NHT results among researchers. One of the most problematic logical flaws is the practice of viewing rejection of the null as support for the researcher's non-null research hypothesis. This problem can be viewed as one of researchers testing inappropriate hypotheses. As long as this is the case, no amount of reform can hope to redeem NHT or indeed statistical practice in psychology.

The critics vary in what they wish to see happen regarding statistical practice in psychology. Some call for an outright ban on the use of NHT (e.g., Hunter, 1997), others support a continued, possibly less central use of NHT with the reporting of additional quantities (e.g., Mulaik, Raju, & Harshman, 1997). The APA Task Force (Wilkinson et al, 1999) produced recommendations consistent with reforming rather than replacing NHT. The Task Force made several recommendations regarding both research and statistical practice. The APA has since adopted those recommendations (APA, 2001) in its publications manual. Those relevant to this discussion include:

- increased emphasis on a priori power estimates,
- the reporting of confidence intervals, and
- the reporting of observed effect size measures.

To varying degrees these along with other recommendations all represent positive steps towards the reform of statistical practice in psychology. At least some of these recommendations are already being incorporated into undergraduate statistics texts for psychology students. So far the trend among textbooks incorporating these recommendations is to keep the same general

orientation towards NHT while supplementing it with new or expanded coverage of power analysis, confidence intervals, and effect sizes to meet the Task Force's recommendations.

Critics of NHT believe that these reforms will encourage more thoughtful treatment of data and statistical results than is currently evidenced (Krantz, 1999). While these reforms may indeed help they do not address the fundamental problem that testing a-theoretical hypotheses is largely unproductive no matter what modifications are made to the testing procedure.

NHT IN PSYCHOLOGY

Current statistical training in psychology and some other disciplines (e.g., education) identifies the null hypothesis as a "straw person" hypothesis that is rejected in order to "prove" the research hypothesis. There is no need for the null to be theoretically meaningful in any sense other than it predicts something that the research hypothesis does not. The justification for such hypotheses is that the researcher wishes to have a hypothesis that can be rejected in order to "prove" the research hypothesis. Since all research hypotheses in psychology posit an effect of some, usually unknown, magnitude a null hypothesis of "no effect", a nil hypothesis, is a ready-made "opposite" to the research hypotheses. So much so that for most psychological researchers the nil hypothesis is *the* null hypothesis.

There are three major problems with this approach:

- The null is intended as an a-priori false (straw-person) hypothesis. Its falsehood, and therefore its rejection, is never truly in doubt. The only question is whether the researcher has collected enough data to demonstrate this predetermined result. The rejection of such a null provides no new information and therefore permits no progress.
- The rejection of any null does not "prove" the research hypothesis in any but the weakest sense.
- The practice of rejecting "no effect" in favor of "some effect" does not encourage the researcher to determine the size and nature of the effect.

The problem with the straw-person hypothesis is that it is never intended to represent a value that is even remotely plausible for the parameter being tested. For example the researcher might test the null that men and women do not differ on a psychological characteristic where the differences are already well established (e.g., moral reasoning, aggressiveness, spatial reasoning). In such cases rejection of the null provides absolutely no new information for the researcher, they have merely re-established what was already known.

In this context "an a-priori false hypothesis" refers to a hypothesis that is already well established by prior evidence as being false; a hypothesis that does not even represent a close approximation to the correct value for some parameter. The nil hypothesis has been decried as being "always false" by some (e.g., Cohen, 1994; Meehl, 1967, 1997; Tukey, 1991) making it the embodiment of the a-priori false null hypothesis. Some have argued that there are at least some cases where the nil hypothesis cannot be declared false a-priori and in fact may be true (e.g., Frick, 1995; Wainer, 1999). However such cases seem to be the minority.

As already mentioned the problem with the rejection of a-theoretical nulls is that they provide no new information. In some cases they can even provide misleading information. Consider the case in which a researcher wished to establish reliability or convergent validity for a measure of some psychological characteristic. By convention both of these require that appropriate correlation coefficients exceed .7. However it does happen that the researcher tests the nil hypothesis $\rho = 0$ and takes the rejection of that hypothesis as support for the reliability or convergent validity of the measure!

Popper's (1968/1959) notion of falsification is often incorrectly invoked to support this approach. Rejecting the straw-person hypothesis does nothing more than demonstrate that which was already known, that the straw-person hypothesis was false. Popper argued that theories are corroborated when the researcher challenges the theory, tries to falsify it, and fails. Popper further argues that the falsification of a particular theory does not provide corroboration for any other theory.

Thus the second problem I identified earlier: rejection of any null provides only the weakest possible support for any alternative hypothesis. Meehl (1967, 1997) has discussed this

issue at length so I will only summarize the portion germane to this discussion: the rejection of the nil hypothesis provides support for *all* theories that predict an effect. For example all theories that predict a difference between men and women on measures of aggressiveness will be supported by rejection of the nil hypothesis that no gender difference exists. It does not matter if these theories are based on physiology, social influences, or the position of the constellation Libra, they all gain support from the rejection of the nil. Cohen (1994), a critic of NHT, summarized this issue saying "What I and my ilk decry is the 'weak' form [of NHT] in which theories are 'confirmed' by rejecting null hypotheses" (p. 999).

Finally, the practice of rejecting nil hypotheses never allows the researcher to establish anything beyond the direction of effects. It is not clear if too many psychological theories are too weak to predict anything more than the direction of an effect or if psychological researchers have just learned bad habits from their statistical training. No doubt some of the inability to predict the magnitude of effects results from psychology being a relatively young science. But training which emphasizes searching exclusively for "some effect" rather than a specific effect has to have contributed to this problem. Much of the training received by psychological researchers implicitly discourages making specific theoretical predictions. These researchers are never trained to estimate the values of unknown parameters that would be consistent with their theories.

DIRECTION FOR CHANGE

To address these problems statistical training in psychology must be revised so that

- The null represents a theoretically meaningful value (Fisher, 1955).
- We must stop treating rejection of the null as evidence favoring some specific alternative hypothesis but rather treat it as evidence against the null.
- Start training researchers to specify or estimate parameter values for their theoretical models.

The changes in statistical practice adopted by the APA attempt to address the problems identified through the reforms mentioned earlier. These reforms represent a shift in emphasis towards parameter estimation. However, these reforms do not go far enough. To change statistical practice in psychology it is first and foremost necessary to find some means of encouraging the formulation of theoretically meaningful nulls. It is tempting to recommend teaching NHT with an emphasis on theoretically meaningful nulls. However, when discussing this approach with colleagues I find it difficult to persuade them that null hypotheses can be anything other than nil hypotheses that are rejected in order to "prove" the research hypothesis. That world-view is too deeply ingrained to change easily. An apparently non-NHT approach is required.

I suggest that we "give up" NHT in favor of using a goodness-of-fit approach to the testing of models and model parameters. This is, of course, a sleight of hand change since the difference between NHT, as currently practiced and goodness-of-fit testing (GFT) is that GFT uses theoretically meaningful model based nulls. Thus the reform consists not of banning NHT but rather reframing it into a more appropriate framework. A shift in training and practice to model development and model testing would address the weaknesses inherent in NHT as currently practiced in psychology and education and improves on the Task Force's recommendations.

In many ways such a shift requires no more than a change in the language we use when we teach. Instead of teaching students about "descriptive statistics" we teach "parameter estimation" and the development of "descriptive models of behavior". Instead of learning about "inferential statistics" students would learn about "testing model parameters utilizing goodness-of-fit procedures." GFT is already used extensively by some psychologists (c.f., *Mathematical Psychology*). It is not widely used and is viewed by many as being different from NHT. Because of the widespread lack of familiarity with GFT it is a viable "replacement" for NHT. A replacement that seems different because of its emphasis on good fit but familiar because it produces F-, χ^2-, and p-values. This mix of the new and the familiar should make GFT an easy change to accept.

JUSTIFICATION OF CHANGE

In order to generate theoretically meaningful nulls students need to understand how models are developed, the purpose they serve, and how parameters are estimated in order to appreciate the reasons for testing model parameters. Students need to be introduced to the idea of models and parameter estimation from the very beginning of their training so that they have a framework which allows them to understand which parameter tests (nulls) are meaningful and which are not.

Relying exclusively on confidence intervals and effect sizes there is no reason to believe that NHT practice will actually change in any substantial way. Researchers will still test nil-nulls in order to "prove" their research hypotheses. The computation of confidence intervals and effect sizes may make this practice seem less appealing, but it offers no alternative. In order to make a meaningful change in practice an emphasis on models and GFT is needed.

In some areas of psychology students learn outside of their statistics courses that in a well designed experiment the null hypothesis represents the absence of some parameter from a theoretical model. And they learn that experiments are only useful when the existence of the parameter is in doubt. With an emphasis on models, parameter estimation, and GFT this behavior easily generalizes to testing specific non-zero parameter values. When we test theoretically meaningful nulls we make scientific progress.

As mentioned earlier the rejection of a null hypothesis serves only as evidence against that null. This is difficult to see when years of training and practice have convinced us that rejection of the null 'proves' some unspecified alternative theory. It is easier to recognize that tests of the null provide information about the null when a model is tested for its goodness-of-fit. If the model fits well it is seen as supported and if it does not fit well it is seen as evidence of a poor model. A GFT approach makes it fairly obvious that tests of the null provide evidence about the theory generating the null, not some other theory.

Finally the model building and testing framework makes the estimation of theoretical parameters a natural part of the research process. The estimation of parameters is largely absent from research where the statistical paradigm emphasizes testing to see if the parameter is not equal to zero. While the computation of confidence intervals and effect sizes can encourage the use of more carefully specified models, model building and testing encourages it to a much greater level.

Parameter estimation can certainly be, and in some cases will need to be, accomplished through post-hoc confidence intervals, regression, or other parameter estimation techniques. However, post-hoc estimation never really allows for strong tests of parameter estimation. Unless the researcher has some notion of theoretically rational values of a parameter they will not know if its statistical estimate is reasonable (Roberts & Prasher, 2000). It is simply not possible to know if your model is reasonable when $b = 1.3$ if all you have specified in your model is that $\beta > 0$. On the other hand if the theory predicts that $1.2 \leq \beta \leq 1.5$ then the estimate from the data is clearly consistent with the theory.

Model building and model testing addresses the concerns of NHT critics in a manner that adding confidence intervals and effect size estimates to the common test of nulls cannot. Model testing provides a framework within which the nulls tested are expected to be theoretically meaningful. It discourages the proof of untested theories based on the rejection of tested theories. It encourages researcher to think about their theoretical models more carefully as model testing requires specific parameter values to test.

Since model testing is just a more useful form of NHT there is no need to throw the NHT baby out with the bathwater. This baby can be saved.

REFERENCES

Boring, E.G. (1919). Mathematical versus statistical importance. *Psychological Bulletin, 16*, 335-338.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49, (12)*, 997-1003.

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

Frick, R.W. (1995). Accepting the null hypothesis. *Memory and Cognition, 23*, *(1),* 132-138.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H., (Eds.) (1997). *What if there were no Significance Tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Hunter, J.E. (1997). Needed: A ban on the significance test. *Psychological Science, 8 (1)*, 3-7.

Krantz, D.H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 40*, 448, 1372-1381.

Meehl, P.E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34 (2),* 103-115.

Meehl, P.E. (1997). The problem is epistemology, not statistics. Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In Harlow, L.L., Mulaik, S.A., and Steiger, J.H. (Eds.). *What if there were no Significance Tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Mulaik, S.A., Raju, N.S., & Harshman, R.A. (1997). There is a time and place for significance testing. In Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (Eds.), *What if there were no Significance Tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, *(2)*, 241-301.

Popper, K.R. (1968/1959). *The logic of scientific discovery*. New York: Harper & Rowe.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, *(2),* 358-367.

Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57 (5),* 416-428.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, *(2),* 115-129.

Tukey, J.W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, *(1),* 100-116.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4*, *(2)* 212-213.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychological journals. *American Psychologist, 54,* 594-604.