HOW SIGNIFICANCE TESTS SHOULD BE PRESENTED TO AVOID THE TYPICAL MISINTERPRETATIONS

Stefan Krauss and Christoph Wassner,
Max Planck Institute for Human Development
Germany

*The use of significance tests in science has been debated from the invention of these tests until the present time. Apart from theoretical critiques on their appropriateness for evaluating scientific hypotheses, significance tests also receive criticism for inviting misinterpretations. Although these misinterpretations are well documented, until now there has been little research on pedagogical methods to remove them. Rather, they are considered "hard facts" that are impervious to correction. We discuss the roots of these misinterpretations and propose a pedagogical concept to teach significance tests, which involves explaining the meaning of statistical significance in an appropriate way.*The present contribution is based on Krauss and Wassner (2001) and Haller and Krauss (in press).

INTRODUCTION

The current debate about null hypothesis significance testing (often referred to as NHST) reminds us of a struggle that ends in an impasse between the critics and the defenders. The widely reported criticisms of NHST address – among other issues – its weight in social science, its mathematical principle, its ease of misinterpretation and its mindless use (for a review on the "significance test debate" see, for instance, Nickerson, 2000). At present, both parties seem to have won: On the one hand, the critics of NHST because much of their critique is substantial and largely uncontradicted (Carver, 1993) and on the other, the defenders because in almost all scientific areas (e.g., in all social sciences) NHST is still taught to students as *the* method for evaluating scientific hypotheses.

Our contribution does *not* comment on this debate. Rather, we take the teaching of NHST as a given fact and focus on *improving* it. Unfortunately, literature suggests that after a statistics course the average student cannot describe the underlying idea of NHST (e.g., Falk and Greenbaum, 1995) What is mastered is the mere calculation of a significance test. Yet, in our view the teaching of NHST can only be justified if students are able to grasp the *meaning* of what they are doing.

The commonly accepted view seems to be that students in the social sciences – who are not interested in statistics – show a natural confusion about an inherently difficult concept that invites misunderstandings regardless of the way it is taught. The aim of this contribution is twofold: In the first section, we question this view: Do instructors of methodology (e.g., people who are responsible for teaching NHST to psychology students) address and clarify the meaning of NHST in their lectures? Are they themselves aware of the correct interpretation of a significant test result at all? Although the existence of misinterpretations of significance – at least among students – can be considered a well-known fact, there is astoundingly little pedagogical effort to eliminate these misconceptions. In the second section, we provide a pedagogical concept of how to teach significance testing that is explicitly designed to do away with the typical misinterpretations.

AN EXPERIMANTAL STUDY ON THE UNDERSTANDING OF SIGNIFICANCE

In 6 German universities, we asked participants from psychology departments if they would fill out a short questionnaire. We chose psychologists because in Germany studying psychology entails more statistical lectures than studying other social sciences. We sorted them into three groups: The group *methodology instructors* ($N = 30$) consisted of university teachers who taught psychological methods including statistics and NHST to psychology students. In Germany, some of these teachers are scientific staff (including professors who work in the area of methodology and statistics), and some are advanced students who teach statistics to beginners (so called "tutors"). The group *scientific psychologists* ($N = 39$) consisted of professors and other scientific staff who are *not involved* in the teaching of statistics. The last group consisted of *psychology students* ($N = 44$).

To all three groups, we presented a questionnaire consisting of six statements representing common illusions of the meaning of a significant test result. This questionnaire is an adaptation of Oakes (1986) who tested 68 academic psychologists (however, in Oakes' version, the hint that "several or none of the statements may be correct" was not included).

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means t-test and your result is (t = 2.7, d.f. = 18, p = 0.01). Please mark each of the statements below as "true" or "false". "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

1) *You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).* [ ] true / false [ ]
2) *You have found the probability of the null hypothesis being true.* [ ] true / false [ ]
3) *You have absolutely proved your experimental hypothesis (that there is a difference between the population means).* [ ] true / false [ ]
4) *You can deduce the probability of the experimental hypothesis being true.* [ ] true / false [ ]
5) *You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.* [ ] true / false [ ]
6) *You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.* [ ] true / false [ ]

The correct interpretation (a *p-value* is the probability of the available - or of even less likely - data, given that the null hypothesis is true) is not among the statements. Before presenting our results, let us see why all six statements are wrong: Statements 1 and 3 are easily classified as being false: Significance tests can never *prove* (or *disprove*) hypotheses. Significance tests provide probabilistic information and can, therefore, at best be used to corroborate theories. Statements 2 and 4 should be classified as false because it is generally impossible to assign a probability to any hypothesis by applying significance tests: One can neither assign it a probability of 1 (statements 1 and 3) nor any other probability (statements 2 and 4). Making statements about probabilities of hypotheses is only possible in the alternative approach of Bayesian statistics. We will later pick up on this approach as a basic element of our pedagogical concept on how to clarify what a significant test result does mean and what it does *not* mean. Statement 5 may look similar to the definition of an error of Type I (i.e., the probability of rejecting the H_0 although it is in fact true), but having actually rejected the H_0 (as in statement 5), this decision would be wrong, if and only if the H_0 were true. Thus the probability in statement 5 ("...that you are making the wrong decision") is $p(H_0)$, and this probability – as we learned from statement 2 – cannot be derived with NHST. Statement 6 reflects the so-called "replication fallacy". In Neyman and Pearson's paradigm, one could interpret $\alpha = 0.01$ in a frequentistic framework as relative frequency of rejections of H_0 *if H_0 is true*. The example however gives no evidence of the H_0 being true. "In the minds of many, 1 - p erroneously turned into the relative frequency of rejections of H_0, that is, into the probability that significant results could be replicated" (Gigerenzer, 1993).

RESULTS

The percentage of participants in each group who erroneously classified at least one of the statements as correct is shown in Figure 1. Oakes' (1986) original findings are displayed on the far right. Astonishingly, nearly 90 % of the *scientific psychologists* perceive at least one of the false "meanings" of a p-value as true. However, our finding that even among the *methodology instructors* 80% share these misinterpretations is flabbergasting. Since it can be assumed that the topic of "significance" is addressed frequently during their lectures, this fact is difficult to believe. As expected, the performance of the methodology instructors is nevertheless somewhat better than the performance of the other scientists, while the performance of psychology students is the worst. This order is also reflected in the mean number of wrongly endorsed statements: It was 1.9 for the methodology instructors, 2.0 for the other scientists, and 2.5 for the psychology students..In any case, we can summarize: 4 out of every 5 *methodology instructors* have misconceptions about the concept of significance, just like their students.

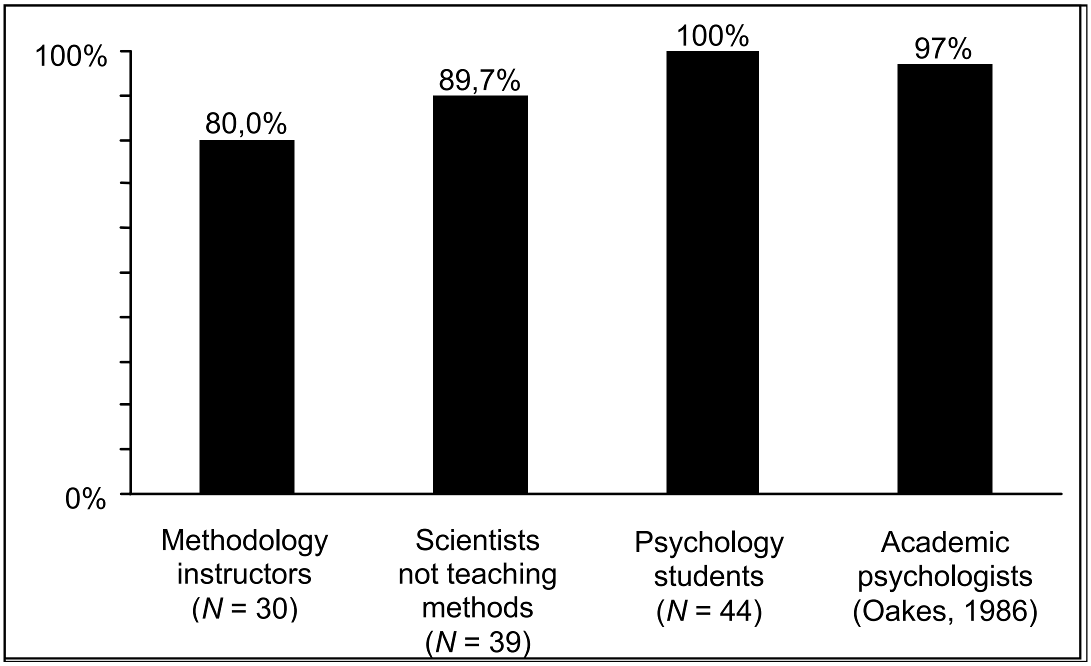

Figure 1. Percentage of Participants in Each Group, Who Erroneously Classified At Least One of the Statements as Correct.

What can we do to counteract the belief in these six statements? Falk and Greenbaum (1995) report that it is not much help to simply point out the misinterpretations to students. They suggest “that unless strong measures in teaching statistics are taken, the chances of overcoming this misconception appear low at present” (p. 93). In the following we provide a pedagogical concept that helps make clear to students what a significant test result *does* mean, what it does *not* mean and, *why* not. The main aim of our pedagogical approach is to *contrast NHST with Bayes’ rule* by highlighting the differences between both approaches: If one wants to prevent students belief that NHST says something about the probability of hypotheses, one should explicate the approach that *actually* can. Unfortunately, in textbooks – as well as in students’ minds – both statistical approaches are completely separate. If at all, Bayes’ rule is taught just as a formula without mentioning that it is the basis for an alternative approach of hypothesis testing. Often, even the symbols for data (D) and hypotheses (H) are not used in the formula, but instead just two arbitrary events A and B are represented. In order to establish a link between both approaches, the idea of NHST should be expressed - just like Bayes' rule - in terms of conditional probabilities, and Bayes’ rule should be taught including the concepts of hypothesis (H) and data (D). Thus, our pedagogical concept includes teaching the following steps.

FOUR STEPS TOWARDS AN UNDERSTANDING OF THE MEANING OF NHST

Step one: Teach students that there are two statistical inference paradigms based on conditional probabilities: To check the “fit” between our data (D) and a hypothesis (H), probability theory provides two approaches relying on conditional probabilities, namely NHST and Bayesian Statistics.

Step two. Teach students the underlying idea of NHST: Considering p(D | H): The probability of data (D) given a hypothesis (H) is assessed in significance testing: In Fisher’s paradigm, D represents the presently available data (or data that are even less likely) and H represents the null hypothesis. Expressing the result of NHST in terms of conditional probabilities – namely as $p(\mathrm{D}\,|\,\mathrm{H}_0)$ – makes two facts salient: First, only statements concerning the probability of data can be obtained, and, second, H_0 functions as a given fact. This latter issue indicates that any paraphrasing of a significant test result must refer to this fact, e.g.: "*... given H_0 is true*".

Step three. Teach students the Bayesian inference approach: Considering p(H | D): To find out the probability of a hypothesis (H) given data (D), we can apply Bayes' rule:

$$p(\mathrm{H}\,|\,\mathrm{D}) = \frac{p(\mathrm{D}\,|\,\mathrm{H})\cdot p(\mathrm{H})}{p(\mathrm{D}\,|\,\mathrm{H})\cdot p(\mathrm{H}) + p(\mathrm{D}\,|\,\neg\mathrm{H})\cdot p(\neg\mathrm{H})}$$

Bayesian statistics, which is based on this formula, is the only methodology that allows us to make statements about the (conditional) probabilities of hypotheses. The probability of hypotheses is what we have in mind when testing hypotheses. Our claim is that only when students learn the meaning of $p(\mathrm{H}\,|\,\mathrm{D})$ as derived from Bayes' rule, will they no longer believe that this probability is what results from NHST. Rather, they will understand this probability as being *inverse* to the concept of *p*-values.

Step four. Insight by comparison: Now we are ready to elucidate misinterpretations concerning NHST by contrasting the antonyms:

(a) $p(\mathrm{D}\,|\,\mathrm{H}_0)$ is what you derive from NHST.

(b) $p(\mathrm{H}\,|\,\mathrm{D})$ can be obtained by Bayesian Statistics.

Our claim is that presenting these two propositions will facilitate the understanding of the crucial essence of NHST. Note that presenting the underlying idea of Bayesian statistics – assessing $p(\mathrm{H}\,|\,\mathrm{D})$ – here is sufficient. In depth instruction in Bayesian statistics is not required. Of course, also the paradigm of Neyman and Pearson can be expressed in this way and added to the list:

(c) $p(\mathrm{D}\,|\,\mathrm{H}_0)$ and $p(\mathrm{D}\,|\,\mathrm{H}_1)$ are taken into account with H_1 being the experimental hypothesis.

Comparison of paradigms (a) and (c), by the way, counteracts the problematic "hybridisation" of NHST (according to Fisher) and hypothesis testing including a specific alternative hypothesis (according to Neyman and Pearson). Gigerenzer (1993) used the term "hybridisation" to point out that despite many conflicts, both paradigms are mixed in most statistics text books. This practice blurs the crucial underlying differences. Our didactical lineup clearly separates both approaches: Whereas in (a) the probability of data is just considered in *one* possible "world" (namely H_0), in (c) two possible "worlds" are explicitly taken into account (H_0 and H_1). Having internalized these four steps, how would a student now react to the six statements? Statements 1-5 all belong to the Bayesian world. Furthermore, as we have already learned, the notion of conditional probabilities reveals that NHST always refers to a "world" in which the H_0 is true. Because this is not a trivial statistical matter of course, without this specification a description of a result of NHST can never be correct. Since this requirement is reflected in *none* of the statements, all six statements (including statement 6) must be wrong.

REFERENCES

Carver, R.P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287-292.

Falk, R., & Greenbaum, W. (1995). Significance tests die hard. *Theory & Psychology, 5*, 75-98.

Gigerenzer, G. (1993). The Superego, the ego, and the id in statistical reasoning. In G. Keren and C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Erlbaum.

Haller, H., & Krauss, S. (in press). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online.*

Krauss, S., & Wassner, C. (2001). Wie man das Testen von Hypothesen einführen sollte. *Stochastik in der Schule, 21, 1*, S. 29-34.

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

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chichester: Wiley.