Methods of Psychological Research Online 2002, Vol.7, No.1 Institute for Science Education
Internet: http://www.mpr-online.de

Misinterpretations of Significance: A Problem Students Share with Their Teachers?

Heiko Haller[1] & Stefan Krauss[2]

Abstract

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. We presented six common misinterpretations to psychologists who work in German universities and found out that they are still surprisingly widespread – even among instructors who teach statistics to psychology students. Although these misinterpretations are well documented among students, 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.

1. 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

[1] Free University of Berlin, Dept. of Educational Science and Psychology, Otto-von-Simson-Straße 19, 14195 Berlin; E-Mail: mail@heikohaller.de

[2] Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin;
E-Mail: krauss@mpib-berlin.mpg.de

because much of their critique is substantial and largely uncontradicted (Sedlmeier, 1996; Carver, 1993) and on the other, the defenders because NHST is still taught to psychology students at universities as *the* method for evaluating scientific hypotheses.

This article does *not* comment on this debate (for recent contributions concerning the debate on NHST in this journal see Sedlmeier, 1996; Iseler, 1997; Sedlmeier, 1998; Hager, 2000). 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 avarage student cannot describe the underlying idea of NHST (Falk and Greenbaum, 1995; Gigerenzer and Krauss, 2001). 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 lack of understanding NHST can be demonstrated by asking students the following question: "What does the statement that a statistical measure is significant at the 5% level *mean*?" It is not very likely to receive the correct answer, namely: "The probability of the available (or of even less likely) data, given that the null hypothesis is true, is less than 5%." Tragically, this answer is not just an answer to an inconsequential statistical question. Rather, this answer represents the understanding of the underlying idea of the whole inference approach.

An overview of students' misinterpretations of this question reveals that there are basically two main classes: The first class contains "meaningless answers" like: "It means, that the measure lies 5% above the random-percentage". The second class of misinterpretations contains the belief that with NHST the probability of hypotheses can be assessed.[3] Furthermore, there are individual cases deviating from these classes, such as the statement that significance says something about the replicability of a significant test result. It is important to note that most students showing such misinterpretations have mastered the *formal procedure* of significance testing.

The aim of this article is twofold: In the first section, we describe possible sources of the widespread confusion. The commonly accepted view seems to be that psychology students – who are not interested in statistics – show a natural confusion about an inherently difficult concept that invites misunderstandings regardless of the way it is

[3] For a collection of wrong beliefs stemming from a written undergraduate examination (Vordiplom) in psychology see Gigerenzer and Krauss (2001, pp. 53-54).

taught. In the first section, our questions are: Do instructors of methodology (i.e., people who are responsible for teaching NHST to psychology students) address and clarify the meaning of NHST in their lectures? Furthermore: Are they aware of the correct interpretation of a significant test result at all? Our empirical results reveal that clarifying the *meaning* of NHST to psychology students in Germany is usually not a matter of statistics education. We found that most methodology instructors do not even know the correct interpretation but rather share the misconceptions of their students.

Although this lack of insight – 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.

Where do the Misconceptions come from?

As previous experimental studies show, almost all students (Falk & Greenbaum, 1995) and even most academic psychologists (Oakes, 1986) have no real insight into the meaning of a significant test result. Does this confusion principally arise in students' minds or is it supported – or even implemented – from outside? For the implementation assumption, we basically have two suspects, namely (1) statistical textbooks or (2) statistics teachers.

Indeed textbooks have already been detected as a possible source of misconceptions. An especially striking example is the book by Nunally (1975) *Introduction to statistics for psychology and education*. Within three pages (pp. 194-196), he provides the following eight interpretations of a significant test result that all are wrong:

- "the improbability of observed results being due to error"
- "the probability that an observed difference is real"
- "if the probability is low, the null hypothesis is improbable"
- "the statistical confidence ... with odds of 95 out of 100 that the observed difference will hold up in investigations"
- "the degree to which experimental results are taken 'seriously'"
- "the danger of accepting a statistical result as real when it is actually due only to error"
- "the degree of faith that can be placed in the reality of the finding"

- "the investigator can have 95 percent confidence that the sample mean actually differs from the population mean"

Finally, Nunally (1975) concludes: "All of these are different ways to say the same thing". Even if this amount of explicit teaching of confusion is an exception, almost all statistical textbooks put more weight on the execution of formal procedures than on the understanding of the meaning of the results. Taking into account that these books are written by psychologists, this is an astounding fact. Although a warning like "a significant test result can tell us nothing about the probability of H_0 or H_1 being true" could easily contribute to clarification, it is missing in most textbooks. Sedlmeier and Gigerenzer (1989, p. 314) provide references of authors as well as editors of journals who explicitely give wrong interpretatons of significance.[4] Be that as it may, things could be straightened out in university lectures: Our emprirical research addresses the question of whether the second group of suspects, the statistics teachers, remove misunderstandings or contribute further to the confusion about NHST.

2. Survey

2.1. Method

In 6 German universities, we asked participants from psychology departments if they would fill out a short questionnaire. 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* ($N = 44$).

[4] Of course there are also many competent statistics text books. An excellent example, for instance, is Hays (1963).

To all three groups, we presented the following questionnaire consisting of six statements representing common illusions of the meaning of a significant test result:[5]

2.2. The Questionnaire

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 [ ]

In the introduction, we already provided the real meaning of a statistical test result that is significant at the 5% level. A *p-value* consequently is the probability of the available (or of even less likely) data, given that the null hypothesis is true. Before presenting our results, let us see why all six statements are wrong.

[5] Our questionnaire was a german adaptation of Oakes' (1986) who tested 68 academic psychologists. In Oakes' version, the hint that "several or none of the statements may be correct" was not included.

Statements 1 and 3

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, 4, and 5

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). Therefore, statements 2 and 4 should be classified as false. 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.

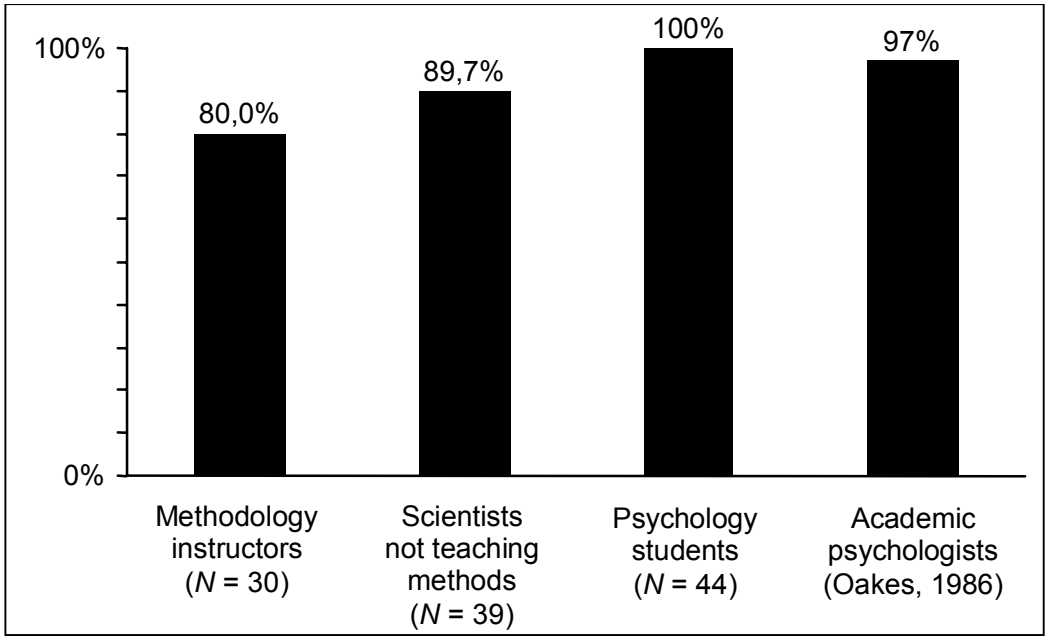
Statement 6

Statement 6 reflects the so-called "replication fallacy". In Neyman and Pearson's paradigm, one could interpret $\alpha = .01$ 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, 1993a). A well-known example of this fallacy came from the former editor of the Journal of Experimental Psychology, A.W. Melton (1962, p. 553): "The level of significance measures the confidence that the results of the experiment would be repeatable under the conditions described".

2.3. 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.

Figure 1: Percentages of participants in each group who made at least one mistake, in comparison to Oakes' original study (1986).



Astonishingly, despite the fact that 15 years have passed since the publication of Oakes' book and despite publication of numerous articles on the misunderstandings connected with significance testing, nothing seems to have changed. Nearly 90% of the *scientific psychologists* perceive at least one of the false "meanings" of a p-value as true. However, our novel finding that even among the *methodology instructors* 80% share these misinterpretations is flabbergasting.[6] Since it can be assumed that the topic of "significance" is addressed frequently during their lectures, this fact is difficult to believe.

Group comparisons reveal that – as expected – the performance of the methodology instructors is nevertheless somewhat better than the performance of the other scientists,

[6] The make up of the *methodology instructors group* ($N = 30$) roughly reflects the relationship in which statistics in Germany is actually taught to psychology students: Apart from 3 professors, we had in this sample 17 scientists and 10 advanced students ("tutors") who teach statistics. This relationship of statistics teachers (3-6 scientists and tutors per professor) is not unusual for larger German universites. It is, however, important to note that the 6 methodology instructors (= 20%) who correctly de-masked all six statements as being wrong included all three *professors* examined.

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: Teaching statistics to psychology students in German universities does not include effective enlightenment about the *meaning* of significance: Four out of every five *methodology instructors* share misconceptions about this concept with their students.

Let us consider each statement separately. Which statements were especially hard to de-mask, and which were easy to see through? Table 1 shows our findings broken down into single statements.

Table 1: Percentages of different wrong answers (i.e., statements marked as true) in the various groups and in comparison to Oakes' original study (1986)

	Germany 2000 Psychological Departments of German Universities			USA 1986 (Oakes)
Statements (abbreviated)	Methodology Instructors	Scientific Psychologists (*exept* Methodology)	Psychology Students	Academic Psychologists
1) H_0 is absolutely disproved	10	15	34	1
2) Probability of H_0 is found	17	26	32	36
3) H_1 is absolutely proved	10	13	20	6
4) Probability of H_1 is found	33	33	59	66
5) Probability of Type I error	73	67	68	86
6) Probability of replication	37	49	41	60

Across all groups, the first three statements were easier to de-mask than the last three. Although statements 2 and 4 must either both be wrong or both be true,[7] they were not checked consistently, but statement 4 was endorsed more frequently. This might be due to wishful thinking because in psychological experiments H_1 is usually the more interesting hypothesis. Furthermore, Table 1 reveals that statement 5 is the most popular one.

What can we do to counteract the belief in these six statements? In the following section 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.

3. How can we Avoid Misconceptions in Students' Minds?

Falk and Greenbaum (1995) report that it is not much help to simply point out the misinterpretations to students. In a study, they recruited 53 students and used statements similar to Oakes'. Although their students read Bakan's (1966) article, which explicitly warns against wrong conclusions, and although the right alternative ("None of the statements is correct") was among their statements, only 13% of their participants opted for the right alternative. Falk and Greenbaum (1995) suggest "that unless strong measures in teaching statistics are taken, the chances of overcoming this misconception appear low at present" (p. 93).

The main aim of the pedagogical approach we suggest 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

[7] Because $p(H_1) = 1 - p(H_0)$, knowing the probability of one hypothesis also means knowing the probability of the complementary hypothesis.

concepts of hypothesis (H) and data (D). Thus, our pedagogical concept includes teaching the following steps:

3.1. Four Steps Towards an Understanding of the Meaning of NHST

3.1.1. 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.

3.1.2. 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.[8] Expressing the result of NHST in terms of conditional probabilities – namely as $p(D \mid H_0)$ – 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*".

3.1.3. Step three: Teach students the underlying idea of 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:

$$p(H \mid D) = \frac{p(D \mid H) \cdot p(H)}{p(D \mid H) \cdot p(H) + p(D \mid \neg H) \cdot p(\neg H)}$$

[8] This $p(D \mid H_0)$ is exactly the p-value we get from NHST: the probability of the available (or of even less likely) data, given that the null hypothesis is true. It is important to note that the notation of a *p*-value as $p(D \mid H_0)$ only makes sense if 1) H_0 is interpreted as an event and if 2) H_0 is a point hypothesis: If H_0 were a composite hypothesis, we would have to calculate the maximum of the probability of rejection across all parameters in the parameter space. Only if H_0 were specified with exactly this parameter, one could derive a *p*-value, which then admittedly could not be expressed as $p(D \mid H_0)$.

Bayesian statistics, which is based on this formula, is the only methodology that allows us to make statements about the (conditional) probabilities of hypotheses.[9] The probability of hypotheses is what we have in mind when testing hypotheses.[10] Our claim is that only when students learn the meaning of $p(\mathrm{H} \mid \mathrm{D})$ 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.

3.1.4. Step four: **Insight by comparison**

Now we are ready to elucidate misinterpretations concerning NHST by contrasting the antonyms:

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

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

The contrasting 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 (1993a) 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. This "hybridisation" of two – intrinsically incompatible – inference approaches can

---

[9] For a detailed discussion of the Bayesian inference approach see, for instance, Edwards, Lindman and Savage (1963) or Howson and Urbach (1989).

[10] Regarding the discrepancy between the outcome of NHST and researchers' actual interest, Cohen (1994, S. 997) made the following famous remark: "What's wrong with significance testing? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe in that it does! What we want to know is 'given these data, what is the probability that H_0 is true?' But as most of us know, what it tells us is 'given that H_0 is true, what is the probability of these (or more extreme) data?'."

be considered another potential cause of students' confusion. 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.[11] 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.

3.2. Supporting measures

The four steps proposed on how to overcome an 'amazing persistence of a probabilistic misconception' (Falk & Greenbaum, 1995) should be accompanied by measures that inform students about the background of NHST. In addition to the correct *paraphrasing*, this will guarantee a well-informed *evaluation* of a significant test result. Therefore, teaching NHST to psychology students should dispel the following three myths that can be observed in students' thinking (taken from Gigerenzer and Krauss, 2001):

3.2.1. Myth 1: The level of significance is a well-defined concept

The confusion of $p(D \mid H)$ with $p(H \mid D)$ is not the only statistical fog in students' minds. An additional problem often is making the clear distinction between the level of significance α and a p-value. We propose presenting the historical development of these concepts. Deliberate comparing of conflicting approaches might not only result in a clear distinction of the differences, but, furthermore, can give statistics an interesting face. The following table can animate students to think thoroughly about the term "level of significance":[12]

Table 2 reveals that confusion about the level of significance is not something to be ashamed about: Even Sir Ronald Fisher changed his mind concerning this concept. Ta-

[11] In the first statement $p(H \mid D) = 0$ and in the third statement $p(H \mid D) = 1$.

[12] The "whole story" on the struggle accompanying the invention of significance tests can be found in Gigerenzer et al. (1989).

ble 2, furthermore, allows us to associate ideas with names rather than just with definitions.

Table 2: Interpretations of "level of significance"

"Early" Fisher (1935)	"Late" Fisher (1956)	Neyman and Pearson
The level of significance has to be determined *before* conducting a test (in the sense of a convention, e.g. $\alpha = 5\%$). Thus, the level of significance is a property of the *test*.	The exact level of significance has to be calculated *after* a test is conducted (p-value). Here, the level of significance is a property of the *data*. An arbitrarily determined convention is no longer required.	α and β have to be determined *before* conducting a test. α and β are the relative frequencies of an error of the first or second kind and are therefore properties of the *test*. Yet, to determine α and β no convention is required, but rather a cost-benefit estimation of the severity of the two kinds of error.

3.2.2. Myth 2: NHST are scientifically taboo

NHST are usually taught as if they were undisputed as, for instance, geometrical theorems. Yet, presenting NHST as an uncontroversial method is not telling the whole truth. Indeed, the end of the debate about NHST cannot be foreseen. Without telling students about this ongoing struggle, one cannot blame them for believing that NHST is established like geometry. The following articles – collected from the last ten years – are a good way of showing students that statistics is *not* geometry:

Table 3: Significance testing is not geometry

Loftus, G.R. (1991):	On the tyranny of hypothesis testing in the social sciences.
Gigerenzer, G. (1993b):	Über den mechanischen Umgang mit statistischen Methoden.
Cohen, J. (1994):	The earth is round ($p < .05$).
Loftus, G.R. (1994):	Why psychology will never be a real science until we change the way we analyze data.
Dar, R., Serlin, D., & Omer, H. (1994):	Misuse of statistical tests in three decades of psychotherapy research.
Falk, R., & Greenbaum, W. (1995):	Significance tests die hard.

After reading one or more of these articles, students will interpret results of NHST in a more careful way and will consider NHST to be what it is: A mathematical tool that can weaken or corroborate scientific hypotheses. Students will appreciate that NHST is *no panacea* which per se fits every problem and – most importantly – that it *cannot* replace statistical thinking. As suggested by their titles, many articles criticising NHST are highly emotional. As a teacher, it is important to remain as neutral as possible when introducing students to this debate. Meehl (1978), as a counter example, condemned NHST unequivocally: "I believe that the almost universal reliance on merely refuting the null hypothesis ... is ... one of the worst things that ever happened in the history of psychology" (p. 817). Adopting such statements as one's own is of little help in improving students' statistical understanding.

Nevertheless, introducing the debate on NHST into teaching is not only indispensable if we want to teach the truth, it furthermore has motivational advantages. Gigerenzer (1993a, p. 327) writes: "Some statistics teachers protest that raising these disputes would only confuse students. I believe that pointing out these conflicting views would make statistics much more interesting to students who enjoy thinking rather than being told what to do next."

This view, by the way, fits pedagogical theories well. Berlyne (1965), for instance, dicussed the concept of "epistemic curiosity". He hypothesized that epistemic curiosity results from conceptual conflict, as is the case when new information appears to contradict earlier understandings (see also Slavin, 1994, p. 370).

From experience in seminars, we can report that students indeed do participate in the "scientific fight" on hypothesis testing with a great deal of attention and curiosity.

3.2.3. Myth 3: NHST is the only way to test scientific hypotheses

The counter example of myth 3 is already included in our pedagogical concept that required the introduction of Bayes' rule. The US-American National Council of Teachers of Mathematics (NCTM) in 2000 published new guidelines for improving the teaching of mathematics that generally propagate pluralism of methods – not only in the field of statistics. According to this NCTM-Standards 2000, students should be able to "apply a wide variety of strategies" and "to select and use various types of reasoning and proof as appropriate". Teachers should take care to "discuss various approaches to solve problems". A collection of alternative methods is given by Sedlmeier (1996) in this journal. This collection is particularly appropriate, since it contains a lot of intuitive methods –

like error-bars, effect-sizes or various kinds of explorative data analysis (e.g., box-plots). Unlike NHST or advanced Bayesian methods, most of these methods can be taught with little effort and are easy to understand.

Asked why they perform NHST, most students would probably answer: "Because it is the rule". It is important to note that *statistical thinking*, which is promoted in this article, is not based on *which* statistical method is applied, or, whether any method is applied at all. Of course, all mentioned statistical methods can be applied routinely by just following a recipe. Sound statistical thinking includes beyond it the ability to *justify* the respective procedure – whichever was chosen. Let us take as an example for this claim the present article.

Following a suggestion from Loftus (1993) that sometimes "a picture is worth more than a thousand p-values" we presented the relevant results pictorially: In the present article it is possible to convey all relevant messages *without* conducting hypothesis testing. The rationale is the following: Our Survey was conducted to assess the actual performance of the respective groups. It was not questioned whether methodology instructors will perform better than students; this was rather the default assumption. The interesting point was whether and to what extent methodology instructors exhibit misconceptions. The result that 80% in this group had misconceptions is remarkable – regardless of the performance of other participants. Even if only 50% of the methodology instructors showed misconceptions, this result would be alarming. Thus, assessing confidence intervals is dispensable. We think that *everyone* who teaches statistics should know the meaning of the most frequent word in this subject.[13]

4. Conclusion

NHST procedures are designed in such a way that allows them to be carried out without understanding the underlying principle: Just calculate according to a recipe and then reject or accept H_0. So, do we need an intervention at all to help students' understanding of the concept of significance? The two physicians Beck-Bornholdt and Dubben (1998) analyzed the use of significance tests in scientific medical journals. Under the

[13] Currently, in the DFG-Project *Verbesserung des Stochastikunterrichts in der Schule* the present proposals are tested in intervention programs at German secondary schools with pupils of the 12th and 13th grade. Afterwards, interventions with psychology students will follow and we hope to present these results in one of the next issues of this journal.

title 'The unbearable lightness of significance', they claimed that misunderstanding the meaning of a significant test result in almost all cases means according 'significance' *more* importance than it deserves: Sometimes, "significance" is even interpreted as a mathematical proof, making considering further evidence unnecessary. As a consequence, medication is often introduced that is indeed ineffective or, the reverse effect, a promising new medication is rated as *definitely* ineffective because an experiment yields a p-value of .06.

Of course, medical researchers are not the only ones who do not seem to be aware of the arbitrariness of the value $\alpha = 5\%$. The American Association of Psychology (APA) wrote in their Publication Manual 1974: "Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such" (S. 19). Most students would probably espouse this statement which considers NHST as surrogate for statistical thinking. A more enlightened view was provided by Rosnow and Rosenthal (1989): "Surely, God loves the .06 nearly as much as the .05".[14]

Regarding NHST, when is *statistical thinking* required? Performing statistical calculations by following a recipe is not what we understand by statistical thinking: For this purpose we have computer software. Statistical thinking is required at the beginning – that is, choosing the appropriate inference method – and at the end – that is, interpreting the result of any statistical inference process. The present pedagogical approach of comparing different inference paradigms targets at improving the interpretation of a test result. In order to achieve an improvement of students' ability to choose the appropriate method, it is crucial to present fundamentally different problems that require different solutions.

When teaching various inference approaches, it is important not to fall for the trap of dogmatism and just to *replace* NHST by another method. There is no "best" statistical inference method per se. Gigerenzer (1993a) points out that "it is our duty to inform our students about the many good roads to statistical inference that exist, and to teach them how to use informed judgment to decide which one to follow for a particular problem" (p. 335). In this journal Iseler (1997) warns "a knife is hardly appropriate to eat

[14] Indeed Fisher seems to have determined 5% just because he could not get permission to reprint the tables of his rival Karl Pearson (the father of Egon Pearson) that also included other levels of significance (Gigerenzer, 1998).

soup with and a spoon is difficult to cut with. Luckily, that we have both – we should not play off one against the other".

To sum up: Teaching statistics to psychology students should not only consist of teaching calculations, procedures and formulas, but should focus much more on statistical thinking and understanding of the methods. Although it can be assumed that this is an implicit aim of lectures on statistics, our survey suggests that these efforts are not sufficient. Since *thinking* as well as *understanding* are genuine matters of scientific psychology, it is astounding that these issues have been largely neglected in the methodology instruction of psychology students.

Acknowledgements

We want to thank Laura Martignon, Kathryn B. Laskey, Callia Piperides, Anita Todd, Silke Atmaca, Matthias Licha and two anonymous reviewers for reading the manuscript and giving helpful comments.

References

[1] Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437.

[2] Beck-Bornholdt, H.P. & Dubben, H.H. (1998). *Der Hund, der Eier legt. Erkennen von Fehlinformationen durch Querdenken.* Rowohlt Taschenbuch Verlag Reinbek bei Hamburg.

[3] Berlyne, D.E. (1965). Curiosity and Education. In: J.D. Krumboltz (Ed.), *Learning and the Educational Process.* Chicago: Rand McNally.

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

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

[6] Dar, R., Serlin, D., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*, 75-82.

[7] Edwards, W., Lindman, H., & and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242.

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

[9] Fisher, R.A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

[10] Fisher, R.A. (1956). *Statistical methods and scientific inference.* Edinburgh: Oliver and Boyd.

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

[12] Gigerenzer, G. (1993b). Über den mechanischen Umgang mit statistischen Methoden. In E. Roth (Ed.), *Sozialwissenschaftliche Methoden.* Munich, Germany: R. Oldenbourg Verlag.

[13] Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences, 21*, p. 199-200.

[14] Gigerenzer, G. & Krauss, S. (2001). Statistisches Denken oder statistische Rituale? Was sollte man unterrichten? In: Borovcnik, M., Engel, J. & Wickmann, D. (Hrsg.), *Anregungen zum Stochastikunterricht: Die NCTM-Standards 2000, Klassische und Bayessche Sichtweise im Vergleich.* Franzbecker: Hildesheim, S. 53-62.

[15] Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: how probability changed science and everyday life.* Cambridge: Cambridge University Press.

[16] Hager, W. (2000). About some misconceptions and the discontent with statistical tests in psychology. *Methods of Psychological Research Online, 5, 1*, p. 1-31.

[17] Hays, W. L. (1963). *Statistics for Psychologists.* New Nork: Holt, Rinehart and Winston.

[18] Howson, C. & Urbach, P. (1989). *Scientific Reasoning. The Bayesian Approach.* Open Court Publishing Company.

[19] Iseler, A. (1997). Signifikanztests: Ritual, guter Brauch und gute Gründe. Diskussion von Sedlmeier, 1996. *Methods of Psychological Research Online.*

[20] Loftus, G.R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary psychology, 36*, Vol. 36, 2, p.102-105.

[21] Loftus, G.R. (1993): A picture is more worth than thousand p-values: On the irrelevance of hypothesis testing in the computer age. Behavior Research Methods, Instrumentation and Computers*, 25*, 250-256.

[22] Loftus, G.R. (1994). *Why psychology will never be a real science until we change the way we analyze data.* Paper presented at Psychonomic Society meetings, Washington DC.

[23] Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

[24] Melton, A.W. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553-557.

[25] NCTM - National Council of Teachers of Mathematics (2000). *Principles and Standards for school mathematics.* Reston, Va.: National Council of Teachers of Mathematics.

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

[27] Nunally, J.C. (1975). *Introduction to statistics for psychology and education.* New York: McGraw-Hill.

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

[29] Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

[30] Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research Online, Vol. 1, No. 4*, S. 41-63.

[31] Sedlmeier, P (1998). Was sind die guten Gründe für Signifikanztests? *Methods of Psychological Research Online, Vol. 3, No. 1*, S. 39-42.

[32] Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309-316.

[33] Slavin, R.E. (1994). *Educational Psychology. Theory and Practice.* Allyn and Bacon: Boston.