Kühberger et al. BMC Research Notes (2015) 8:84
DOI 10.1186/s13104-015-1020-4

BMC Research Notes

RESEARCH ARTICLE Open Access

The significance fallacy in inferential statistics

Anton Kühberger[1*], Astrid Fritz[2], Eva Lermer[3] and Thomas Scherndl[1]

Abstract

Background: Statistical significance is an important concept in empirical science. However the meaning of the term varies widely. We investigate into the intuitive understanding of the notion of significance.

Methods: We described the results of two different experiments published in a major psychological journal to a sample of students of psychology, labeling the findings as 'significant' versus 'non-significant.' Participants were asked to estimate the effect sizes and sample sizes of the original studies.

Results: Labeling the results of a study as significant was associated with estimations of a big effect, but was largely unrelated to sample size. Similarly, non-significant results were estimated as near zero in effect size.

Conclusions: After considerable training in statistics, students largely equate statistical significance with medium to large effect sizes, rather than with large sample sizes. The data show that students assume that statistical significance is due to real effects, rather than to 'statistical tricks' (e.g., increasing sample size).

Keywords: Statistical significance, Practical significance, Effect size, NHST, Sample size

Background

There is continuing debate on the usefulness and validity of the method of Null Hypothesis Significance Testing (NHST, e.g., [1-3]). Several journals edited special issues on this topic (e.g., *Journal of Experimental Education* in 1993; *Psychological Science* in 1997; *Research in the Schools* in 1998) that culminated in the question: What is beyond the significance test ritual (*Journal of Psychology* in 2009)?

The debate has led to an increased awareness of the problems associated with NHST, and these problems are linked to what has been referred to as a 'crisis of confidence' [4]. Among the dominant recommendations for NHST is reporting of effect size as a supplement to the *p* value [5]. Accordingly, not only the statistical significance of a result should be valued but also the effect size of the study (e.g., [1,6-12]). This should prevent readers from holding the false belief that significant results are automatically big and important, or otherwise, that not significant means 'no effect at all'. Although these misconceptions, that significance means big, and non-significance means no effect, are often referred to in the literature (e.g., [3,13-17]) their empirical basis is weak. This is clearly in conflict with the demand for evidence based practice in statistics and statistics education [18]. Thus, the purpose of the present study was to investigate the prevalence of these misconceptions.

Statistical and practical significance

The distinction between statistical and practical significance is quite old. The origin of statistical significance can be traced back to the 1700s [19]. Practical significance, expressed as the strength of the relationship between two variables, can roughly be dated back to the 18th century [20]. Modern statistical significance refers to the *p* value as the result of a significance test. If $p < .05$ a result is statistically significant. This notion of statistical significance became popular in the social sciences in the first half of the 20th century mainly due to the work of Sir Ronald Fisher [21,22]. With the rise of the statistical significance test, the concept of effect magnitude became seemingly dispensable. Only recently, there is an opposite trend and many authors pointed to the importance of reporting the magnitude of the effect under investigation, mostly because statistical tests are so heavily influenced by sample size (e.g. [6,23-32]). Recall that a test statistic is the product of sample size and effect size [16,33]. The *p* value, as a common-language translation of the various test statistics [8], is therefore also a function of practical significance and sample size, in short: $p = f(ES, N)$. If the effect is small

* Correspondence: Anton.kuehberger@sbg.ac.at
[1]Department of Psychology and Centre of Cognitive Neuroscience, University of Salzburg, Hellbrunnerstr. 34, 5020 Salzburg, Austria
Full list of author information is available at the end of the article

but the sample size very large, the p value will be statistically significant. Similarly, if the effect size is large and the sample size small, the p value will also be significant. Thus, given a big enough sample, even trivial effects can be statistically significant [34]. A correct interpretation of the significance test therefore requires taking the relationship between sample size and effect size into account [35].

Consider the classic aspirin textbook example in Rosnow and Rosenthal [36]: A study tested the effect of aspirin on reducing heart attacks. 11.034 men were given an aspirin pill to be taken every 2 days, whereas 11.037 other men were given a placebo. Statistically speaking, the treatment was enormously effective ($p < .000001$). It was so effective that it was decided to end the study prematurely because the outcome was clear and it appeared unethical to deprive the participants of the control group of the beneficial aspirin [37]. Here statistical significance was equated with practical significance. However, the treatment was far from being effective in terms of effect size ($r^2 = .0011$): statistical and practical significance can tell different stories.

The failure to distinguish between statistical and practical significance has been called the *significance fallacy* [17]. It comes in two varieties. The first variety is to equate a low p value with a big effect size. Thus, the numeric value of p is considered as an indicator of the strength of the treatment effect under test, or the distance between the groups that are compared. Kline [16] called this the *magnitude fallacy*. The second variety of not distinguishing between statistical and practical significance is that statistically non-significant results are interpreted as evidence of no effect, as 'no difference between means', or as 'no relationship between variables'. We call this the *nullification fallacy*.

The nullification fallacy has the potential to damage science (and lives), as an example taken from Fidler (Fidler F: From statistical significance to effect size estimation: statistical reform in psychology, medicine and ecology. Unpublished PhD thesis, University of Melbourne, 2006) shows: in ecology, mark-and-recapture studies are used to determine population sizes. To identify individual frogs upon recapture researchers used to clip certain combinations of toes in order to spare the sensitive skin of the animals, since some studies investigating whether toe-clipping had an impact on the frog's survival rates found no significant effects. But when Parris and McCarthy [38] reanalysed the evidence they found that toe-clipping did actually decrease the survival rate by 6–18% with each toe clipped. The sample sizes of the original studies were just too low to (statistically) show the effect. In this example, the consequence of misinterpreting a non-significant result as indicating that there is no effect is obvious. Other researchers have heatedly argued about the negative consequences of uncritically using underpowered studies to declare the null hypothesis true (for animal studies see [39]; for Neuroscience, see [40]).

Previous research

It is difficult to estimate how common these fallacies about statistical and practical significance are, although these misconceptions are discussed in virtually every article reviewing NHST (e.g. [3,13-17]). Among the first experimenters to examine the level of $p < .05$ were Rosenthal and Gaito [41,42], who found an abrupt drop in confidence in a p level just above .05. Nelson et al. [43] later found this *cliff effect* in a survey on psychological researchers. Again and again the magical nature of $p < .05$ has been shown (e.g., [44]), with some authors proposing strategies for adjusting effect size estimates taking the publication bias into account (e.g., [45,46]). More recent research investigated the consequences of the cliff effect with an eye on the unhealthy effect of small (i.e., underpowered) - however significant – studies for psychological research in general (e.g., [47,48]).

Among the range of studies discussing the consequences of NHST, two empirical studies investigated intuitions about the relationship between effect size, sample size and p. Oakes [49] asked academic psychologists to estimate the unstandardized effect size for a given Student's t-test example. Oakes prescribed the p value and sample size and had participants estimate the size of the effect (see Figure 1). He found that the effect size was generally overestimated. In addition, for identical sample sizes, participants understood that the effect size associated with a p value of .01 is bigger than the effect size associated with a p value of .05. The increase in effect size for $p = .01$ is normatively correct, but participants overstated it considerably. That is, as the p decreased the effect size was assumed to increase disproportionally. This can be seen as an instance of the *magnitude fallacy*.

Wilkerson and Olson [35] investigated how graduate students understand the relationship between effect size, sample size, and errors of statistical inference by asking which of two studies, one with a small and one with a big sample size, provides better evidence when both are statistically significant at $p = .05$. Only one out of 52 graduate students recognized that, given two different studies reporting the same p value, the study with the smaller sample size indicates the larger effect. Missing the link between sample size and effect size is also indicated in the legendary hospital problem by Kahneman and Tversky [50], where participants failed to recognize that a smaller hospital is more likely to have an uneven portion of male vs. female babies born than a large hospital.

In Figure 1 the studies of Oakes [49] and Wilkerson and Olson [35] are schematically depicted, and the main outcomes are reported. As can be seen, each study focused on a specific part of the relationship between effect

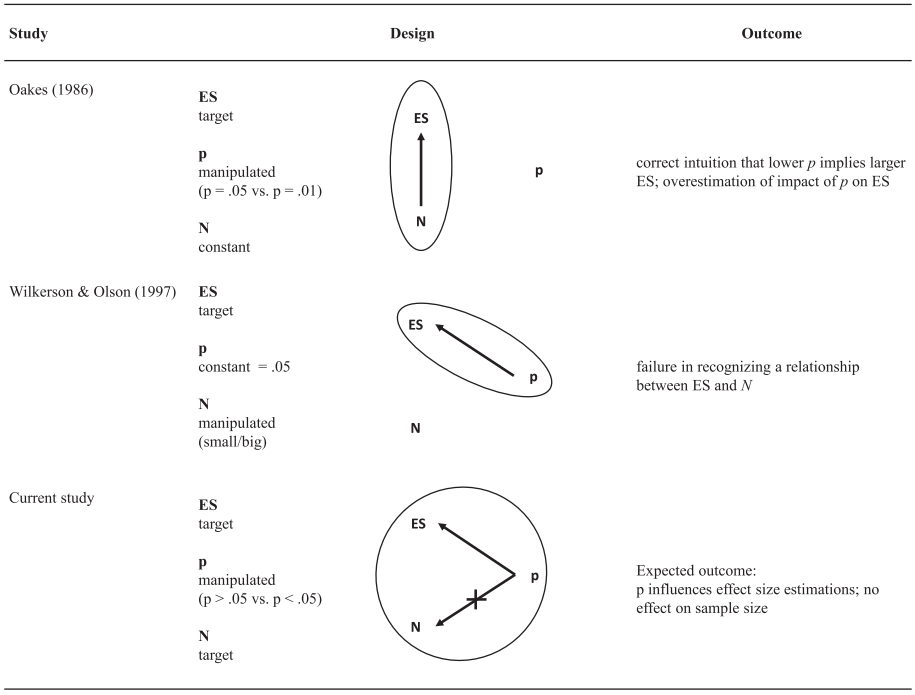

Figure 1 Overview of previous studies investigating the understanding of the relationship between effect size (ES), and sample size (N).

size, sample size and *p* (circled in Figure 1), holding one parameter constant. Here we extended this design to include all three variables concurrently for understanding the relationship between effect size, sample size and *p* inside the heads of our participants.

The present research

In the present research we examined whether students associate statistical significance with practical significance, as stated by the *significance fallacy*. We did this by testing both aspects of the *significance fallacy*: the *magnitude fallacy* and the *nullification fallacy*. According to the *magnitude fallacy* statistically significant results will be rated as having a higher effect size than non-significant results; according to the *nullification fallacy,* non-significant results will be rated as zero in effect size. In essence, we investigated intuitions about effect sizes and sample sizes in the context of *p*-values: is a significant *p* value due to effect size or due to sample size?

Our procedure was as follows: we picked two published studies and described the procedure and the aim of these studies, including the main hypothesis and the dependent measure. We also described whether or not the finding was statistically significant. Note that we did not report statistical measures such as sample size, means, or standard deviations; we rather had participants estimate these measures. That is, we reported the interpretation of the findings and had participants estimate the data. Statistical inference goes usually the other direction, from data to interpretation. In order to investigate intuitions about data, reversing the inferential direction can lead to important insights. In particular, we compared participants' estimations in the significant vs. the non-significant condition. In this way we were able to test whether people expect the difference between statistical significant and non-significant result in the effect size, the sample size, or both.

Methods

Participants and procedure

We sampled 214 students of psychology (156 females, mean age = 23.5, SD = 6.81) from the University of Salzburg enrolled in a statistics course as participants. Sampling was done in different years, thus the sample is of different cohorts whose statistics education is similar, however. All participants were familiar with hypothesis testing and with the concept of statistical inference due to three previous statistics courses (each 3 hours/week). Students participated during their regularly scheduled class time. To ensure commitment participants were offered the chance of winning 20 Euro. The prize was awarded to the two students who came closest to the actual sample size. Under Austrian law it is not necessary to seek formal ethical approval for conducting this research.

Material and design

We selected two published studies from *Psychological Science:* 'Thermometer of social relations' [51] and 'Body locomotion as regulatory process' Koch et al. [52]. The

'thermometer'-study investigated the influence of different temperatures on social relations. It was tested whether participants rated their social proximity to another person as closer when holding a warm compared to a cold beverage. Proximity was measured on a scale ranging from 1 to 7. The 'locomotion'-study addressed the significance of the motor system in influencing cognitive processes. It tested whether stepping backwards enhances cognitive control, measured by a Stroop test, in comparison to stepping forward. Reaction time in ms was the outcome variable of interest. These two articles were chosen because their main research question and their main outcome variable are easily comprehensible. Both studies used t-tests for the analysis.

Participants were presented a short description of each study of approximately 150 words including research question, method, design (i.e. between groups), and outcome variable (see Appendix) Subsequently, participants were asked to estimate the respective measures as they would expect them to be reported in the results section of the paper (i.e., sample sizes, means and standard deviations of both groups, Cohen's d [53]). All participants were presented both studies consecutively, whereas one was described as statistically significant and the other as non-significant. Note that therefore participants did not rate both, the significant and the non-significant condition of the same study. The sequence of the studies was altered between groups. From the participants' ratings we computed an unstandardized effect size (mean difference, calculated by subtracting the means of the two groups).

Results

An initial survey of the data indicated that some of our participants were unable or unwilling to follow the instruction. To ensure adequate data quality we therefore settled for a rigorous regime of data inclusion. In a first step, in the thermometer-study 13 participants had to be excluded because estimates were beyond the range of the response scale. We then excluded participants with missing values in our main dependent variables ($n_{thermo} = 6$; $n_{locomotion} = 10$), and participants showing signs of inconsistency between p level and condition, either by giving p values larger than 1 ($n_{thermo} = 16$; $n_{locomotion} = 20$), by reporting significant p values in the non-significant condition ($n_{thermo} = 25$; $n_{locomotion} = 23$), or by providing non-significant p values ($p > .05$) in the significant condition ($n_{thermo} = 28$; $n_{locomotion} = 28$). This led to a final sample size of 127 participants in the thermometer scenario, and 133 participants in the locomotion scenario. In terms of power, we achieved a power larger than 0.80 to detect a difference between conditions of $d = .50$, $p < .05$, one-sided test, in both scenarios.

The assumption of normal distribution was violated for the mean difference ratings and for the sample size ratings. In addition, several outliers were included in our data. We therefore computed non-parametric analyses (Mann–Whitney U-Tests) to assess differences between the two conditions. The ratings of sample sizes, median values, and standard deviations, as well as the resultant unstandardized and standardized effect sizes are presented in Tables 1 and 2 for the 'thermometer' and 'locomotion'-study, respectively. For every variable three values are given: (i) the actual result as reported in *Psychological Science*, (ii) the estimations of the participants in the significant condition, and (iii) the estimations of the participants in the non-significant condition. Due to non-normality, we report medians for the two latter conditions. We found neither effects of order of the scenarios ($z < -1.01$, $p > .30$, $r < .09$) nor of the order of significant and non-significant condition ($z < -1.19$, $p > .23$, $r < -.10$) on our dependent variables and thus collapsed these conditions in the further analysis.

Testing the magnitude fallacy

According to the magnitude fallacy, statistically significant results will be rated as having a higher effect size than non-significant results. Therefore we tested whether the absolute unstandardized (M_{diff}), and the standardized effect size (Cohen's d), were higher in the significant than in the non-significant condition. We found that participants

Table 1 Results for 'thermometer'-study

	Actual[a]	'Significant' (n = 53)	'Non-significant' (n = 73)	(z-value) *p-value*	Effect size
N	33	76	50	(z =-1.75) $p = .08$	r = -.15
M_{group1}	5.12	2.70	3.50		
M_{group2}	4.13	4.05	4.00		
M_{diff}	0.99	2.00	1.00	(z =-5.27) $p < .001$	r = -.47
SD_{group1}	1.22	1.00	1.25		
SD_{group2}	1.41	8.00	10.00		
Cohen's *d*	0.78[b]	0.60	0.30	(z =-3.88) $p < .001$	r = -.34

Note. [a]The actual study reported a significant effect. Attempts to replicate the effect of temperature on social relations within the Many Labs Replication Project have failed, however [54].

[b]Cohen's d = .78 is reported in the paper. Calculating effect size from means and standard deviations using the Campbell effect size calculator available at http://www.campbellcollaboration.org/resources/effect_size_input.php results in d = .75, 95% C.I = [0.05; 1.46].

Table 2 Results for 'locomotion'-study

	Actual[a]	'Significant' (n = 65)	'Non-significant' (n = 68)	(z-value) *p-value*	Effect size
N	38	60	50	(z =-0.90) p = .37	r = -.08
M_{group1}	712	150	150		
M_{group2}	676	120	118		
M_{diff}	36	50	10	(z =-2.48) p = .013	r = -.21
SD_{group1}	83	10	5		
SD_{group2}	95	8	5		
Cohen's *d*	0.79	0.70	0.20	(z =-4.16) p < .001	r = -.36

Note: [a] The actual study reported a significant effect.

in the significant condition estimated both effect size measures (M_{diff} and Cohen's *d*) higher than participants in the non-significant condition. These findings were consistent in both studies ('thermometer'-study, for M_{diff}: z =-5.27, p < .001, r = −.47; for d: z =-3.88, p < .001, r = −.34; c.f. Table 1; 'locomotion'-study, for M_{diff}: z =-2.48, p = .013, r = −.21; for d: z =-4.16, p < .001, r = −.36; c.f. Table 2). Note also that the effect size estimates in the significant conditions were quite close to the actual data reported in *Psychological Science.*

Inspection of Tables 1 and 2 shows that participants rated the sample sizes in the significant conditions only slightly larger than in the non-significant conditions. These effects were small (r = −.15, and r = −.08, respectively), and statistically non-significant in both studies. Interestingly, the estimated sample sizes were consistently higher than the sample size of the actual study, which was quite low, however.

Another way to present the results is in terms of ratios of the effect sizes and sample sizes between the significant and non-significant condition of each study. The effect sizes were in both studies rated higher in the significant compared to the non-significant condition (M_{diff}: 2 : 1 and 5 : 1; Cohen's *d*: 2 : 1 and 3.5 : 1, for locomotion and thermometer study, respectively). In contrast, sample sizes were rated only slightly higher in the significant condition in the 'thermometer'-study (1.5: 1) and similarly in the 'locomotion'-study (1.2 : 1).

Testing the nullification fallacy

According to the *nullification fallacy* statistically non-significant findings will be interpreted as evidence of no effect and therefore the effect size should be rated as approximately zero. We found that only a minority of our participants specified the mean difference or Cohen's d as exactly zero (6% in 'thermometer'- study, and 3% in locomotion study, respectively). However, a large proportion of participants thought that an effect in a non-significant study is very small (59% in the 'thermometer'-study, and 60% in the 'locomotion'-study; cf. Figure 2). Table 3 shows the exact distribution in terms of d values: as can be seen, non-significant studies were considered to have very small or small effects, whereas significant studies were estimated as more diverse: very small as well as large effects were estimated. In sum, although most participants did not specify the difference between the two means, or Cohen's d, as exactly zero, they guessed that statistically non-significant findings might also be of low practical significance.

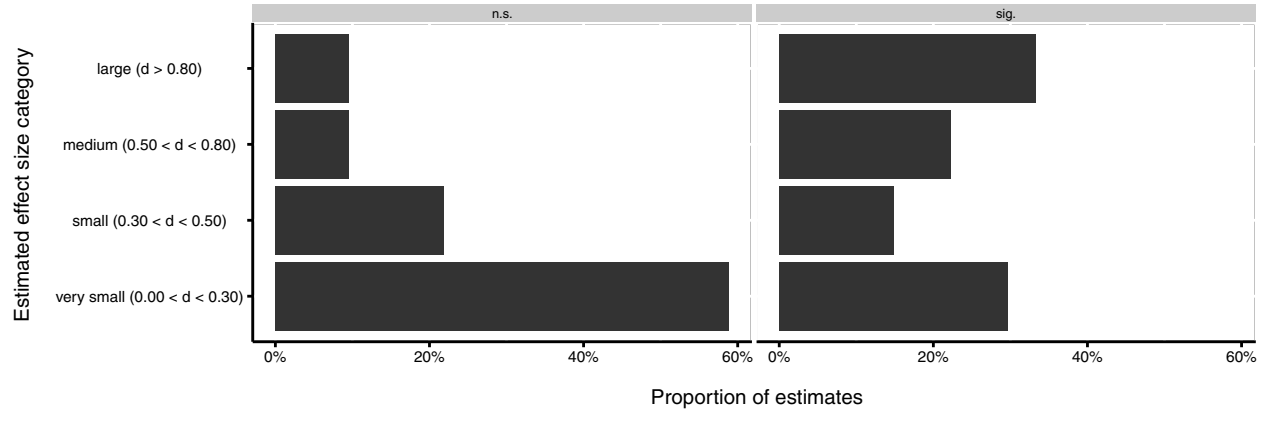

Figure 2 Proportions of participants estimating the effect as very small (0.00 < d < 0.30), small (0.30 < d < 0.50), medium (0.50 < d < 0.80), or large (d > 0.80), for both studies.

Table 3 Crosstabulation of estimated Cohen's d for significant and non-significant condition for both studies

Category	Thermometer study		Locomotion study	
	'Significant' (n = 53)	'Non-significant' (n = 73)	'Significant' (n = 65)	'Non-significant' (n = 68)
Large (d > 0.80)	17 (30%)	7 (9%)	24 (37%)	11 (16%)
Medium (0.50 < d < 0.80)	12 (15%)	7 (9%)	19 (29%)	4 (6%)
Small (0.30 < d < 0.50)	8 (23%)	16 (22%)	13 (20%)	12 (18%)
Very small (0.00 < d < 0.30)	16 (30%)	43 (59%)	9 (13%)	41 (60%)

Discussion

This study tested two fallacies associated with statistical significance: the *magnitude fallacy* and the *nullification fallacy*. According to the *magnitude fallacy* results accompanied by low *p* values are interpreted as having a higher effect size than results with higher *p* values. Effects of non-significant results will, according to the *nullification fallacy,* be interpreted as evidence of no or a negligible effect.

We found that significant results were rated to have higher effect sizes compared to non-significant results. In contrast, sample sizes were not rated higher in the significant condition. That is, in the formula $p = f(ES, N)$ only the effect size seems to be considered, but not the sample size. This could indicate that students assumed that the presented studies have used power analysis to attain the adequate sample size for their experiments. In power analysis the relationship between effect size and samples size is optimized: sample sizes are chosen to be 'big enough' so that an effect of such magnitude as to be of scientific significance will also be statistically significant, but sample sizes will not be 'too big', so that an effect of little scientific importance is not statistically detectable [55,56]. However, power surveys of psychological articles reveal again and again (e.g., [57-63]) that the probability of finding a significant effect of medium effect size (i.e., $r = .30$, $d = .50$) is in the range of 0.40–0.60. This implies that the 'optimal' sample size is rarely calculated. Note also that power analysis is virtually never reported in journal articles as has been shown in a current review [64] that assessed and reanalysed reporting practices of over 6000 educational and psychological articles. This neglect of power is not surprising given that the concept of power arose out of the Neyman-Pearson approach of hypotheses testing which seems to be very seldom used [65].

However, even if researchers (and teachers) fail in using power considerations in their research, it could still be that students rely on power calculations, since this is what they (at least our students) are told in their research methods class. Frequently power considerations have been part of the curricula in the last years, but these ideas seem to meet with little love in current research. In any case, we do not think that considerations of power are the basis for expecting large effect sizes for significant findings.

It is one important feature of our findings that students estimated effect sizes to be different for significant and non-significant findings. Notice that the significant condition was a description of the original study. Thus we can see how close our participants' estimates came to the actual findings. In terms of Cohen's d, participants did not overestimate the effect size, not even in the significant condition. Note, however, that the actual finding of d = 0.75 in the thermometer study is a very imprecise estimate, with the 95% C.I. for d ranging from d = 0.05 to d = 1.46. Hardly any plausible positive effect size estimate can be far off this value. For the locomotion study this presumably also applies, but it is impossible to calculate the exact 95% C.I. from the data due to the within subjects design and the failure to report the correlation between the conditions.

We found only partial support for the nullification fallacy, that non-significant effects will be actually rated as zero in effect size. Only few participants rated the effect in the non-significant condition as exactly zero in size. Practically, this expectation is highly unlikely in the first place as students know that there is actually always a difference in some decimal place between two sample means (cf. the fallacy of soft psychology, [31]). However, although the majority of participants did not predict an effect size of exactly zero, they predicted a lion's share of negligible to small effects in the non-significant condition. A comparison of estimated effect sizes in the significant condition shows a striking difference: here mainly medium to large effect sizes were predicted. People thus do not nullify, they rather minimize. Further studies should investigate the sources of these predictions more thoroughly. For example, it could be important whether or not the research hypothesis was perceived as plausible. Although we have not covered this topic in the current studies, we assume that both research hypotheses (temperature influences perception of social proximity, physical inhibition transfers to cognitive processes) are plausible. Plausibility will surely affect estimations of effect size, beyond significance. Sample size may be less influenced by considerations of plausibility, however.

Estimating statistical values obviously was a very difficult task for participants. Note that we used students in their statistics course as participants, and we incentivized the task. Nevertheless, a good share of our participants

was unwilling to provide plausible estimates. We excluded those from our analysis. However, the remaining participants had also difficulties estimating some values, for instance means, and, most notably, standard deviations. This difficulty was most evident in the locomotion study, where the estimates were far off. As it seems, estimating such values is not part of their training in statistics. In conjunction with a difficult dependent variable like reaction time measured in milliseconds, this may render such a task really difficult. However, getting a grasp on what sample statistics mean, and what their plausible range can be, is important for a thorough understanding of statistical results. Therefore, tasks like ours can be used not only for investigating statistical intuitions, but also for providing training in these intuitions.

Conclusions

This study showed that students have a limited understanding of the underlying concepts of statistical inference. Statistical and practical significance were not distinguished properly. Since some of these students might be future researchers this lack of understanding can have a colossal impact on the whole research field [66]. Indeed, recent analysis of effect size reporting practices found that discrepancies between statistical and practical significances were rarely discussed by the authors of articles [67,68]. But as Kline [16] pointed out, circulating misconceptions like the magnitude fallacy may not be solely the fault of users; rather the logical foundation of contemporary NHST is not entirely consistent. To prevent future confusion about statistical and practical significance effect sizes should be routinely reported as recommended by the major associations in psychology [69] and education [70]. The publication manual ([69], p. 34) states: 'For the reader to appreciate the magnitude or importance of a study's finding, it is almost always necessary to include some measure of effect size in the results section.' But as Henson [71] clarified, a realisation thereof will need continued education and explication. Our findings testify to this conclusion.

Our students tend to interpret the label 'significant' as showing that a study found a nontrivial effect size, rather than that it was large. This is legitimate, and therefore not a fallacy proper. However, in many cases significance is achieved through questionable research practices, among which adaptive sampling (i.e., increasing sample size to achieve significance) is a prominent one [72]. In addition, there is a significant correlation between sample size and effect size in psychological research [73], indicating that significance is often due to large samples, rather than to large effects. Our findings thus show that students still believe in the seriousness of scientific conduct, and that scientific journals are filled with papers that have substance in a practical sense: we must not jeopardize this positive view.

Appendix: Task description and instruction

Dear participant, thank you for taking part in our survey. You will see descriptions of two scientific research papers and we ask you to indicate your personal guess on several features of these studies (sample size, p-value, …). It is important that you give your personal and intuitive estimates.

You must not be shy in delivering your estimates, even if you are not sure at all. We are aware that this may be a difficult task for you – yet, please try.

Task 1: The influence of warmth on social distance

In this study researchers investigated the influence of warmth on social distance. The hypothesis was that warmth leads to social closeness. There were two groups to investigate this hypothesis:

Participants of group 1 held a warm drink in their hand before filling in a questionnaire. Participants of group 2 held a cold drink in their hands before they filled in the same questionnaire. Participants were told to think about a known person and had to estimate their felt closeness to this person. They had to indicate closeness on a scale from 1–7, whereas 1 means 'very close' and 7 means 'very distant'.

The closeness ratings of the participants of group 1 were then compared to the closeness ratings of group 2.

Researchers found a statistically significant [non-significant] effect in this study.

Task 2: The influence of body movement on information processing speed

Previous studies have shown that body movements can influence cognitive processes. For instance, it has been shown that movements like bending an arm for pulling an object nearer go along with diminished cognitive control. Likewise, participants showed more cognitive control during movements pushing away from the body. In this study, the influence of movement of the complete body (stepping forward vs. stepping backward) on speed of information processing was investigated.

The hypothesis was that stepping back leads to more cognitive control, i.e., more capacity. There were two conditions in this study: In the first condition participants were taking four steps forwards, and in the second condition participants were taking four steps backwards. Directly afterwards they worked on a test capturing attention in which their responses were measured in milliseconds. The mean reaction time of the stepping forward-condition was compared to the mean reaction time of the stepping backward-condition.

Researchers found a statistically significant [non-significant] effect in this study.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions
AK made substantial contributions to conception and design, analysis and interpretation of data, and drafting the manuscript. AF made substantial contributions to conception and design, analysis and interpretation of data, and drafting the manuscript. EL made substantial contributions to conception and design, data collection, and analysis and interpretation of data. TS made substantial contributions to conception and design, and analysis and interpretation of data. All authors read and approved the final manuscript.

Acknowledgements
This research was supported by a DOC-fFORTE-fellowship of the Austrian Academy of Sciences to Astrid Fritz.

Author details
[1]Department of Psychology and Centre of Cognitive Neuroscience, University of Salzburg, Hellbrunnerstr. 34, 5020 Salzburg, Austria. [2]Österreichisches Zentrum für Begabtenförderung und Begabungsforschung, Salzburg, Austria. [3]Department of Psychology, University of Munich, Munich, Germany.

Received: 23 September 2014 Accepted: 17 February 2015
Published online: 17 March 2015

References

1. Cumming G. The new statistics: why and how. Psychol Sci. 2014;25:7–29.
2. Dienes Z. Bayesian versus orthodox statistics: which side are you on? Perspect Psychol Sci. 2011;6:274–90.
3. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods. 2000;5:241–301.
4. Pashler H, Wagenmakers EJ. Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? Perspect Psychol Sci. 2012;7:528–30.
5. Ives B. Effect size use in studies of learning disabilities. J Learn Disabil. 2003;36:490–504.
6. Cohen J. Things I have learned (so far). Am Psychol. 1990;45:1304–12.
7. Fan X. Statistical significance and effect size in education research: two sides of a coin. J Educ Res. 2001;94:275–83.
8. Greenwald AG, Gonzalez R, Guthrie DG, Harris RJ. Effect sizes and p values: what should be reported and what should be replicated? Psychophysiol. 1996;33:175–83.
9. Hedges LV. What are effect sizes and why do we need them? Child Dev Perspect. 2008;2:167–71.
10. Kirk RE. Effect magnitude: a different focus. J Stat Plan Inference. 2007;137:1634–46.
11. Thompson B. 'Statistical', 'practical', and 'clinical': How many kinds of significance do counselors need to consider? J Couns Dev. 2002;80:64–71.
12. Vacha-Haase T. Statistical significance should not be considered one of life's guarantees: effect sizes are needed. Educ Psychol Meas. 2001;61:219–24.
13. Castro Sotos AE, Vanhoof S, Van den Noortgate W, Onghena P. Students' misconceptions of statistical inference: a review of the empirical evidence from research on statistics education. Educ Res Rev. 2007;2:98–113.
14. Fidler F, Cumming G, Thomason N, Pannuzzo D, Smith J, Fyffe P, et al. Evaluating the effectiveness of editorial policy to improve statistical practice: the case of the Journal of Consulting and Clinical Psychology. J Consult Clin Psych. 2005;73:136–43.
15. Gliner JA, Leech NL, Morgan GA. Problems with null hypothesis significance testing (NHST): what do the textbooks say? J Exp Educ. 2002;71:83–92.
16. Kline RB. Beyond significance testing: reforming data analysis methods in behavioral research. Washington, DC: American Psychological Association; 2004.
17. Silva-Aycaguer LC, Suarez-Gil P, Fernandez-Somoano A. The null hypothesis significance test in health sciences research (1995–2006): statistical analysis and interpretation. BMC Med Res Methodol. 2010;10:No. 44.
18. Beyth-Marom R, Fidler F, Cumming G. Statistical cognition: towards evidence based practice in statistics and statistics education. Stat Educ Res J. 2008;7:20–39.
19. Hacking I. Logic of statistical iinference. Cambridge: Cambridge University Press; 1965.
20. Stigler SM. The history of statistics. The measurement of uncertainty before 1900. Cambridge, Mass: Belknap Press; 1986.
21. Fisher RA. The Design of experiments, 5th ed. 1951. Edinburgh: Oliver & Boyd; 1935.
22. Fisher RA. Statistical methods and scientific inference. Edinburgh: Oliver and Boyd; 1956.
23. Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing: problems, prevalence, and alternatives. J Wildl Manag. 2000;64:912–23.
24. Bakan D. The test of significance in psychological research. Psychol Bull. 1966;66:423–37.
25. Balluerka N, Gomez J, Hidalgo D. The controversy over null hypothesis significance testing revisited. *Methodology*. Eur J Res Meth Behav Soc Sci. 2005;1:55–70.
26. Berkson J. Some difficulties of interpretation encountered in the application of the chi-square test. J Am Stat Assoc. 1938;33:526–36.
27. Carver RP. The case against statistical significance testing. Harv Educ Rev. 1978;48:378–99.
28. Jones A, Sommerlund N. A critical discussion of null hypothesis significance testing and statistical power analysis within psychological research. Nord Psychol. 2007;59:223–30.
29. Lakens D, Evers ERK. Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies. Perspect Psychol Sci. 2014;9:278–92.
30. Meehl PE. Theory-testing in psychology and physics: a methodological paradox. Philos Sci. 1967;34:103–15.
31. Meehl PE. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. J Consult Clin Psychol. 1978;46:806–34.
32. Thompson B. If statistical significance tests are broken/misused, what practices should supplement or replace them? Theor Psychol. 1999;10:165–81.
33. Rosenthal R. Meta-analytic procedures for social research. 2nd ed. New York: Sage; 1991.
34. Kalinowski P, Fidler F. Interpreting significance: the differences between statistical significance, effect size, and practical importance. Newborn Infant Nurs Rev. 2010;10:50–4.
35. Wilkerson M, Olson MR. Misconceptions about sample size, statistical significance, and treatment effect. J Psychol. 1997;131:627–31.
36. Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. Am Psychol. 1989;44:1276–84.
37. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing physicians' health study. N Engl J Med. 1989;321:129–35.
38. Parris KM, McCarthy MA. Identifying effects of toe clipping on anuran return rates: the importance of statistical power. Amphibia Repilia. 2001;22:275–89.
39. Macleod M. Why animal research needs to improve. Nature. 2011;477:511.
40. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14:365–76.
41. Rosenthal R, Gaito J. The interpretation of levels of significance by psychological researchers. J Psychol. 1963;55:33–8.
42. Rosenthal R, Gaito J. Further evidence for the cliff effect in the interpretation of levels of significance. Psychol Rep. 1964;15:570.
43. Nelson N, Rosenthal R, Rosnow RL. Interpretation of significance levels and effect sizes by psychological researchers. Am Psychol. 1986;41:1299–301.
44. Poitevineau J, Lecoutre B. Interpretation of significance levels by psychological researchers: the .05 cliff effect may be overstated. Psychon Bull Rev. 2001;8:847–50.
45. Bradley MT, Brand A. A correction on the bradley and brand method of estimating effect sizes from published literature. Theor Psychol. 2014;24:860–2.
46. Bradley MT, Stoica G. Diagnosing estimate distortion due to significance testing in literature on detection of deception. Percept Mot Skills. 2004;98:827–39.
47. Bakker M, Wicherts JM. The (mis) reporting of statistical results in psychology journals. Behav Res. 2011;43:666–78.
48. Bakker M, van Dijk A, Wicherts JM. The rules of the game called psychological science. Perspect Psychol Sci. 2012;7:543–54.
49. Oakes M. Statistical inference: a commentary for the social and behavioral sciences. New York: Wiley; 1986.
50. Kahneman D, Tversky A. Subjective probability: a judgment of representativeness. Cogn Psychol. 1972;3:430–54.
51. IJzerman H, Semin G. The thermometer of social relations. Mapping social proximity on temperature. Psychol Sci. 2009;20:1214–20.
52. Koch S, Holland RW, Hengstler M, van Knippenberg A. Body locomotion as regulatory process. stepping backward enhances cognitive control. Psychol Sci. 2009;20:549–50.

53. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. New York, NY: LEA; 1988.
54. Lynott D, Corker KS, Wortman J, Connell L, Donnellan BM, Lucas RE, et al. Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). Soc Psychol. 2014;45:216–22.
55. Lenth RV. Some practical guidelines for effective sample-size determination. Am Stat. 2001;55:187–93.
56. Lenth RV. Statistical power calculations. J Anim Sci. 2007;85:E24–9.
57. Acklin MW, McDowell CJ, Orndoff S. Statistical power and the Rorschach: 1975–1991. J Pers Assess. 1992;59:366–79.
58. Bezeau S, Graves R. Statistical power and effect sizes of clinical neuropsychology research. J Clin Exp Neuropsychol. 2001;23:399–406.
59. Clark-Carter D. The account taken of statistical power in research published in the British Journal of Psychology. Br J Psychol. 1997;88:71–83.
60. Cohen J. The statistical power of abnormal-social psychological research: a review. J Abnorm Soc Psychol. 1962;65:145–53.
61. Kazantzis N. Power to detect homework effects in psychotherapy outcome research. J Consult Clin Psychol. 2000;68:166–70.
62. Rossi JS. Statistical power of psychological research: what have we gained in 20 years? J Consult Clin Psychol. 1990;58:646–56.
63. Sedlmeier P, Gigerenzer G. Do studies of statistical power have an effect on the power of studies? Psychol Bull. 1989;107:309–16.
64. Fritz A, Scherndl T, Kühberger A. A comprehensive review of reporting practices in psychological journals: are effect sizes really enough? Theor Psychol. 2013;23:98–122.
65. Hager W. Vorgehensweise in der deutschsprachigen psychologischen Forschung. Eine Analyse empirischer Arbeiten der Jahre 2001 und 2002. [Procedures in German empirical research – an analysis of some psychological journals of the years 2001 and 2002.]. Psychol Rundsch. 2005;56:191–200.
66. Henson RK, Hull DM, Williams CS. Methodology in our education research culture: toward a stronger collective quantitative proficiency. Educ Res. 2010;39:229–40.
67. Alhija FN, Levy A. Effect size reporting practices in published articles. Educ Psychol Meas. 2009;69:245–65.
68. Sun S, Pan W, Wang LL. A comprehensive review of effect size reporting and interpreting practices in academic journals in Education and Psychology. J Educ Psychol. 2010;102:989–1004.
69. APA (American Psychological Association). Publication manual of the American psychological association. 6th ed. Washington, DC: Author; 2010.
70. American Educational Research Association. Standards on reporting on empirical social science research in AERA publications. Educ Res. 2006;35:33–40.
71. Henson RK. Effect-size measures and meta-analytic thinking in counseling psychology research. Couns Psychol. 2006;34:601–29.
72. John LK, Loewenstein GM, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol Sci. 2012;23:524–32.
73. Kühberger A, Fritz A, Scherndl T. Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. PLoS ONE, 9(9), e105825. doi:10.1371/journal.pone.0105825.