Rethinking Statistical Analysis Methods for CHI

Maurits Kaptein
Eindhoven University of Technology / Philips Research
Den Dolech 2, 5600 MB, Eindhoven, the Netherlands
maurits@mauritskaptein.com
+31(0)6 21262211

Judy Robertson
Computer Science, Heriot-Watt University
Edinburgh Campus, EH14 4AS
Judy.Robertson@hw.ac.uk
+44 (0)131 451 8223

ABSTRACT
CHI researchers typically use a significance testing approach to statistical analysis when testing hypotheses during usability evaluations. However, the appropriateness of this approach is under increasing criticism, with statisticians, economists, and psychologists arguing against the use of routine interpretation of results using "canned" *p* values. Three problems with current practice - the fallacy of the transposed conditional, a neglect of power, and the reluctance to interpret the size of effects - can lead us to build weak theories based on vaguely specified hypothesis, resulting in empirical studies which produce results that are of limited practical or scientific use. Using publicly available data presented at CHI 2010 [19] as an example we address each of the three concerns and *promote consideration of the magnitude and actual importance of effects*, as opposed to statistical significance, as the new criteria for evaluating CHI research.

Author Keywords
Usability evaluation, Bayesian Statistics, Research Methods

ACM Classification Keywords
H.5.2 User interfaces: Evaluation/methodology;

General Terms
Experimentation, Measurement.

INTRODUCTION
A core strength of the CHI community is that members bring together expertise from a range of disciplines "as diverse as user interface design, human factors, computer science, psychology, engineering, graphics and industrial design, entertainment, and telecommunications" [27]. Correspondingly, the community has a rich set of design and evaluation practices at its disposal. An important aspect of training new interaction designers is to teach them how to use different data gathering and analysis techniques "flexibly and in combination to avoid biases which are inherent in any one approach" [28, p.290]. While many approaches to evaluation are valid, it is important that researchers are aware of best practice for any given methodology.

Every so often, it is useful to re-evaluate the standard set of techniques used within an approach and consider whether they provide researchers with the tools they need to answer the questions they are interested in, and whether other techniques would in fact serve the community better. Such a debate is currently taking place in the field of psychology, as demonstrated by a recent special collection of papers within *Perspectives on Psychological Science* [21]. Critics of the traditional statistical inference method of significance testing argue that "it is time for researchers to consider foundational issues in inference" [10, p.274]. Similarly, Wagenmakers, et al. conclude that "experimental psychologists need to change the way they conduct their experiments and analyze their data" [31, p.426] and, in the light of recent positive empirical findings in the theoretically implausible area of extra-sensory perception, argue that the statistical strategies used by psychologists are "too weak, too malleable and offer far too many opportunities for researchers to befuddle themselves and their peers" [30, p.425]. In fact this is a long-standing problem; Cohen noted already in 1994 that such criticisms have been made within psychology for forty years [8].

As a field of study that builds upon statistical methods used by psychologists, usability evaluation is subject to the same criticisms. Indeed, a small number of HCI researchers have identified flaws in experimental design and statistical testing in usability studies. In 1998 Gray and Salzman published an in-depth critique of five well known studies of usability evaluation methods, observing that weaknesses in experimental design (threats to statistical conclusion validity, construct validity, and internal and external validity) call into question the reliability of these findings [14]. More recently, Cairn's survey of inferential statistics in BCS HCI conferences and two leading HCI journals noted common problems in reporting of statistical results; failure to check assumptions about the data required by particular tests, over-testing and using inappropriate tests [4]. Dunlop and Baillie [11] aimed to raise awareness within the sub-field of mobile HCI of problems with statistical analysis techniques such as the use of null hypothesis testing in a binary way to approve results, abusing statistical tests, making illogical arguments as a result of tests, deriving inappropriate conclusions from non-significant results, and confusing the size of p-values with effect sizes.

CHI'12, May 5–10, 2012, Austin, Texas, USA.

In our own examination of the CHI 2011 proceedings we found that 35 (out of 306 total, many of which are not quantitative) papers report results from a *t*-test. Of these 35 papers, six do not report *any* descriptive statistics, *none* report a standardized effect size, and only nine make an attempt to describe or interpret the effect size (the difference in means). *None* of the papers related the magnitude of the effect to previous findings in the literature. We believe this shows that the issues we address in this paper are relevant to the CHI community. In this paper we delve deeper into the criticisms raised by earlier scholars by considering an underlying issue: *should statistical significance testing be the "gold standard" for quantitative empirical work within our field?*

In line with the discussion of Ziliak and McCloskey [32] we argue that there are fundamental flaws associated with sciences that are built primarily on the interpretation of *p*-values. In this paper we focus on three common problems:

1. *The fallacy of the transposed conditional* – Researchers often wrongly interpret *p*-values as the probability of the null-hypothesis being true.
2. *A lack of power* – Researchers often pay little attention to null-results without being aware of the potential of their experimental set-up to reject the null when it is false.
3. *Confusion between p-values and estimates of effects*– Researchers often judge a small *p*-value as indicating a (theoretically or practically) important relationship. This however is not generally correct.

In the remainder of this paper we will first conceptually explain the "traditional" statistics that most of HCI's quantitative results rely on. Next, we will address each of the three errors and, using a running example show (a) how they arise, and (b) how they could be mitigated. Our example is based on publically available simulated data previously published at the CHI 2010 conference [19]. We have chosen this data set to avoid singling out studies by other researchers for criticism, and to make it possible for the interested reader to download the data to study the examples for themselves. The reader is referred to [32] and [9] for a discussion of these and related issues in the fields of economics and psychology.

TRADITIONAL STATISTICS

The traditional approach to statistics within many scientific fields is to use significance testing. In this familiar decision making procedure, the null hypothesis is compared to an alternative hypothesis and one or the other is rejected. The great advantage of this decision-making procedure is that long-term error rates are known, and therefore can be controlled. Researchers can control for both Type I and Type II errors. Type I errors occur when the null hypothesis is rejected when it is actually true and can be controlled by specifying an alpha value (before beginning data collection) which specifies the level of significance under which the null hypothesis will be rejected. Type II errors occur when the null hypothesis is accepted when it is actually false: that is, there is an effect that has not been detected. The proportion of times the null is false but was accepted is called beta. The power of an experiment (1- beta) is the probability of detecting an effect given that the effect really exists in the population. If it is sufficiently unlikely that the observed data was generated by a process that is adequately described by the null hypothesis, then the null is rejected and another, alternative, hypothesis is taken as truth. *Sufficiently unlikely* is in most null hypothesis tests defined in terms of a ratio of signal and sampling error.

To understand the basic idea of most hypothesis testing procedures it is useful to consider the one-sample t-test:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

The *t* value is given by the difference of the sample mean $\bar{x}$ and the population mean u_0 (often zero in this particular case) – the signal – divided by the sample standard deviation over the square root of the number of subjects – the sampling (or standard) error. Higher *t*-values indicate that it's less likely that the sample mean (given the standard deviation and the number of observations) would be observed if indeed in the population mean was equal to u_0. Thus high *t*-values lead to low *p* -values: *the probability of observing the current data given the null hypothesis.*

Low *p*-values – those lower than .05 - would in turn drive most researchers to conclude that the null-hypothesis is *not* true, and thus some alternative hypothesis should be accepted. It is easy to see that high *t*-values (and thus low *p*-values) can be obtained through a combination of a large signal (difference between $\bar{x}$ and u_0), and little sampling error (small *s* and / or high *n*).

Misinterpretations of the p-value

This approach, fiercely promoted by Fisher in the 1930's [9], has become the gold standard in many disciplines including quantitative evaluations in HCI. However, the approach is rather counter-intuitive; many researchers misinterpret the meaning of the *p*-value. To illustrate this point Oakes posed a series of true/false questions regarding the interpretation of *p*-vales to *seventy* experienced researchers and discovered that *only two* had a sound understanding of the underlying concept of significance [25].

So what does a *p*-value actually mean? *"…the p-value is the probability of obtaining the observed value of a sample statistic (such as* t, F, χ^2) or a more extreme value *if the data were generated from a null-hypothesis population and sampled according to the intention of the experimenter"* [22, p.293]. Because *p*-values are based on the idea that probabilities are long run frequencies, they are properties of a collective of events rather than single event. They *do not* give the probability of a hypothesis being true or false for this particular experiment, they only provide a description

of the long term Type I error rate for a class of hypothetical experiments – most of which the researcher has not conducted.

PROBLEM I: THE FALLACY OF THE TRANSPOSED CONDITIONAL

The false interpretation of the *p*-value by most researchers brings up the first problem with null-hypothesis testing. Researchers often interpret the *p*-value to quantify the probability that the null hypothesis is true. Thus, a *p*-value smaller than .05 to large groups of researchers indicates – be it conscious or unconscious – that the probability that the null hypothesis is true (e.g. $\bar{x} = u_0$) is very small.

Under this misinterpretation the *p*-value would quantify $P(H_0|D)$ - the probability that the null hypothesis (H_0) is true, given the data (D) collected in the experiment. However, the correct interpretation of the *p*-value is rather different: it quantifies $P(D|H_0)$ – the probability of the data *given that* H_0 *is true*. Researchers who state that it is very unlikely that the null hypothesis is true based on a low *p*-value are attributing an incorrect meaning to the *p*-value.

It is easy to understand why this misconception is incorrect by the following example: consider the probability of being dead after being lynched, $P(D|L)$. Most would estimate this to be very high, say 0.99. However, the mere observation of a dead person does not lead most people to believe that the corpse was lynched – after all, there are many possible ways to die which don't involve lynching. $P(L|D)$ is (correctly, and we think luckily) estimated to be rather small.

The way forward

There is a well established way to link a conditional probability to its inverse, shown by the Reverend Thomas Bayes in 1763 [3] which enables us to compute $P(H_0|D)$ from $P(D|H_0)$:

$$P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D)}$$

Thus, the probability of the null-hypothesis being true given the data depends on the probability of the data given the null-hypothesis, $P(D|H_0)$, the *prior* probability of the null-hypothesis, $P(H_0)$, and the probability of the data $P(D)$. The prior probability refers to the probability of the hypothesis before the current set of data is collected. It can be seen that for the lynching example the prior probability is so low that the transposed conditional is also low.

The probability of the data, $P(D)$, is often difficult to compute directly. In the discrete case it is given by the sum of the probability of the data given all competing hypotheses. However, in practice its computation is hardly necessary: If a researcher wants to decide between two competing hypotheses $P(D)$ will be the same in the computation of both $P(H_0|D)$ and $P(H_{alt}|D)$ and thus merely acts as a normalizing constant. $P(D|H_0)$ is already provided by our *p*-value, so all we need to compute $P(H_0|D)$ or compare $P(H_0|D)$ to $P(H_{alt}|D)$, is a specification of the prior $P(H_0)$: our prior expectancy of our hypothesis being true. For guidance on setting the value for a prior expectation, consult [26] and [9].

Over the last decades this *Bayesian* approach to computing what we actually want to know, *the probability that our hypothesis are true*, has gained increasing attention among statisticians and researchers. These methods are now gaining ground in fields around us, as evidenced by recommendations to use the so-called Bayesian t-test in cognitive science [22] and medicine [12] .

The Bayesian *t*-test allows researchers to compute the ratio of the likelihood of two competing hypothesis, for example the null hypothesis and an alternative hypothesis [22]. The resulting "Bayes Factor" values of greater than 1 indicate evidence for the null hypothesis, and values less than 1 give support for the alternative hypothesis. Heuristics to judge the strength of support for the null and alternative hypotheses given by a range of Bayes Factor values are listed in [31].

The use of Bayesian statistics rather than traditional statistics has considerable practical ramifications. Wetzels et al. computed Bayes Factors for 855 t-tests published in psychology journals and found that while p-values and Bayes Factors did co-vary (small p-values with large Bayes Factors), the strength of evidence was not calibrated [31]. That is, for studies which reported significant effects (alpha = .05 or .01) 70% of them had Bayes Factors indicating only *anecdotal* evidence in support of the alternative hypothesis.

Changing conventions about the statistical tests accepted within a community is a slow business. However, it is no longer the case that the calculations themselves are difficult. Using packages like {BayesFactorPCL} (for the open source analysis software [R]) researchers can now easily compute Bayes factors[24].

Full Bayesian analysis reaches beyond the Bayesian t-test and provides methods for model comparisons as well as test for different hypothesis. In each case the quantity of interest, P(H|D), informs researchers about the decisions to make based on their collected data. We do not advocate a shift from "canned" p-values to "canned" Bayes Factors – researchers' careful interpretations are still vital- but used appropriately, the Bayesian approach is a solution to the *Fallacy of the Transposed* conditional.

Example: Comparing operating systems

In their paper "Powerful and Consistent Analysis of Likert-Type rating Scales" Kaptein et al. [19] present an example dataset that they use to demonstrate a novel type of analysis. The data is publicly available from http://www.nth-iteration.com/study/statistics/. We will use

this dataset in the remainder of this paper to illustrate the three common problems[1].

This (simulated) dataset describes the potential outcomes of a usability evaluation of two different operating systems. While the original dataset contains usability ratings – answers to the statement *"The system was easy to use"* on a seven-point scale – at two points in time, we focus only on the measurements obtained in the first time point. The data describes the scores of participants on this question after using (a) Windows Vista, or (b) Apple Mac OS-X. The dataset provides the obtained scores for three evaluations with different samples sizes: *N=10*, *N=40*, and *N=200*. We believe this dataset – a straightforward rating presenting a comparison between two conditions (between subjects) for differing sample sizes – provides a good numerical example to illustrate the problems raised in this paper.

Table 1 presents the mean score on the statement for the two different groups for each of the 3 sample sizes reported upon in [19]. The results for the *N=200* case are statistically significant with a difference in means of 0.91 points on the seven-point scale.

	Vista		**OS-X**		**Test results**	
	Mean	**S.D.**	**Mean**	**S.D**	***t-value***	***p***
N=10	4.00	2.16	5.17	1.32	.965	*.383*
N=40	3.86	1.25	4.67	1.85	1.574	*.126*
N=200	3.69	1.57	4.60	1.43	4.281	*<0.001*

Table 1.Overview of the data presented in [16] for time point 1.

The result obtained for *N=40*, a more typical case in many HCI studies, is however not straightforward. The *p*-value of *.126* would lead many to conclude that there is no significant difference between the usability of Windows Vista and that of Apple Mac OS-X based on the obtained ratings. Often, the *observed difference* of 0.81 points would be neglected and the results not further discussed.

Researchers however can compute the actual probability of *H_0 given the data* to further interpret their results. Results from a *Bayesian t-test* give a value of 1.53 for the *N=40* case. This Bayes Factor describes the likelihood of the null-hypothesis compared to an alternative, non-informative, hypothesis. In this case, the result would be interpreted as providing *weak evidence in favor of the null hypothesis* (values between 1 and 3); this evidence would be classified as only *"anecdotal"* by [31]. For the *N=10* case the Bayes Factor is 1.69, leading to a similar conclusion. For *N=200* the Bayes Factor leads to a similar conclusion as the standard t-test: a Bayes Factor smaller than 0.01 provides strong evidence against the null hypothesis.

The advantage of using the Bayesian approach here is that it enables researchers to quantify evidence in favor of the null hypothesis. This is not possible with traditional statistics but is of high importance because it enables us to distinguish between cases where the data is inconclusive (such as the *N=10* and *N=40* cases in our example) and cases where there is strong evidence regarding the null hypothesis (as in our *N=200* example).

PROBLEM II: A LACK OF POWER

The use of *p*-values enables researchers to control Type I errors – or the rejection of H_0 while in fact it is true. However, controlling Type II errors (the failure to reject H_0 when it is false) through calculating the *power* of an experiment appears to be attended to less frequently [7]. The power of a statistical test is the long-term probability that a given test *will find an effect assuming that one exists in the population*. Thus, power indicates whether your experimental setup is capable of detecting the effect that you wish to find. The power is a function of sample size, population effect size and the significance criteria (known as the alpha value, which is set by convention at .05).

The standard accepted power within psychology is .80 [6] which means that there would be *20%* (*1-.80*) chance that the researcher fails to reject the null hypothesis when it is false. Reviews of the psychology literature reveal that the majority of published studies lack power, resulting in a confusing literature with apparently contradictory results [23]. In studies with low power, getting a null result is not particularly informative: *it does not distinguish between the cases where the null is true and where the experimental set-up did not detect the null.*

The way forward

What can researchers do to address lack of power in their studies? Maxwell [23] recommends that power calculations should be performed before the experiment is carried out, and that they should be reported as standard in empirical papers. Cohen [7] gives some heuristics for required sample sizes for eight commonly used statistical tests, given the effect size that is deemed important or sought by the researcher. Consider an example which might occur within usability studies: a researcher is comparing two versions of the same interface with a between subjects design using number of errors as a dependent variable. For analysis using a two tailed independent samples t-test with alpha set at .05, with a power of .80 and attempting to detect a medium sized effect (Cohen's d = .30), the researcher should recruit *176 participants in each group*. Netx to Cohen's heuristics, software packages such as the {pwr} package in [R] [5] can be used for more accurate results, or more complex designs. Power calculations would at least make researchers aware of the problem, but what can be done to increase power if it is found to be low?

The most obvious way to increase power is to increase sample size. Of course, this can be impractical in many fields, including HCI, but there are ways around this. For example, Maxwell [23] suggests that in the field of psychology researchers could gain power by running

[1] The [R] code for all the analysis presented in this paper can be retrieved from the same page.

collaborative multi-site trials in which many research groups conduct the same experiment with manageable numbers of participants and pool their results. Hansen and Collins discuss approaches to increasing power, which do not require an increase in sample size [16]. Although their recommendations are intended for epidemiologists, some are pertinent to HCI such as preventing attrition from studies, increasing the difference between groups by appropriately timing follow-up studies, and reducing variance within groups by using a more homogenous set of participants. They also discuss the virtue of using more reliable and appropriate measurement instruments. For example, in the context of HCI, this suggests the more widespread use of thoroughly validated standard attitudinal scales rather than researchers creating bespoke questionnaires specifically for a new study as is often current practice [2]. Such scales should be sensitive enough to capture differences between groups as advised in [12].

Example continued: Power to reject the null if it is indeed false.

We will illustrate the often surprising lack of power in HCI experiments by following up on the results presented for the *N=40* case comparing the usability ratings of Windows Vista and Apple Mac OS-X that we also used to demonstrate Bayes Factors. Here we compute the power of the difference presented in Table 1. The difference in means is 0.81, and the pooled variance is 2.2. This gives an effect-size (Cohen's *D*) of 0.37 [7]. Given the between subjects design with 20 users each, this gives a *power of 0.14*. The inverse of the power, *1-0.14 = 0.86* is the probability of making a type II error: a failure to reject the null when it is false. Thus, given this experimental setup, and the estimated effect size, a researcher would *not detect an effect this size even if it were actually present in the population in 86 out of a 100 similar experiments.*

This low power for the given effect size and sample size again illustrates the point made in the Bayesian analysis: the evidence in favor of the null is only minor because the chances are good that the experimental set-up will fail to detect an effect.

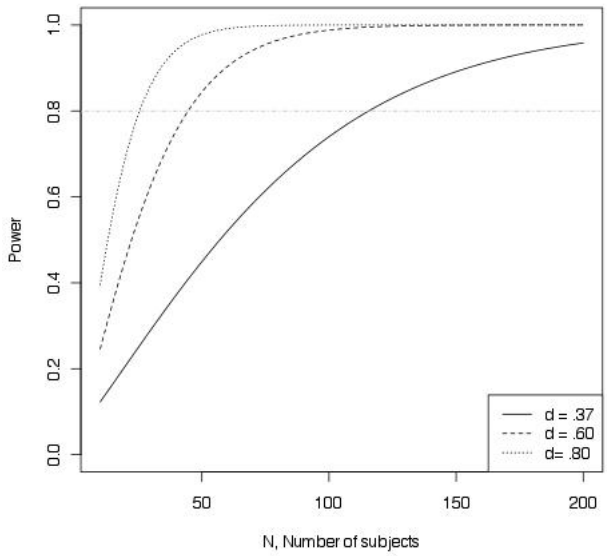

Figure 1.: Overview of power as a function of N, the number of subjects, for different effects sizes in the population.

To provide a better understanding the relationship between effect size, power, and the number of subjects, Figure 1 shows power as a function of *N* for three different effect sizes: 0.37 – the effect size estimated from the data reported in [19] – 0.6, and 0.8. It is clear that both the effect size as well as the number of participants in a study have a large impact on the power of the study. However, it can also be concluded that for small or moderate effect sizes, sample sizes larger than those typical in HCI are necessary.

PROBLEM III: CONFUSION BETWEEN P-VALUES AND ESTIMATES OF EFFECTS

Besides the often-erroneous interpretation of *p*-values and low power, the focus on null hypothesis significance testing in HCI has another severe consequence: qualitative questions about *whether an effect exists* are favored over quantitative questions relating to *how much* of an effect there is, and *to whom it matters.* The latter can only be assessed by considering effect sizes and appropriate loss functions, and by interpreting these in a real world context.

A *p*-value smaller than .05 does not necessarily imply that the effect is important – it only informs us that the sampling error was small compared to the signal. Especially for large data sets (which often lead to powerful tests) low *p*-values are common but do not inform our search for scientific answers. Only a numerical interpretation of the estimated effect can tell us whether a "significant" effect is indeed important to us and warrants further research or a theoretical explanation. For example, if there was a significant difference in the time taken to learn two competing versions of a software package, but the size of the effect was only fifteen seconds, this would likely not have a very large practical impact.

Perhaps surprisingly, the flip side of the argument also holds: a high *p*-value does not imply that the effect under study was unimportant. It only means that it was measured with a relatively high sampling error. Compelling examples of this can be found in neighboring disciplines, and in the courtroom: the painkiller Vioxx was tested in a clinical trial against Naproxen, a general already-on-the-market painkiller. During the trial one person died that was taking naproxen. For Vioxx however, five people died. The difference *was not statistically significant, p >.05,* and thus written off as unimportant. The lawsuits against Vioxx in 2005 proved the researchers wrong: The real-life, and regrettably more powerful, test showed that Vioxx severely – although initially not significantly – raised risks of cardiovascular side effects. If statistical significance is neither a sufficient nor a necessary criterion for importance, what good does it do?

Currently the "size-less stare" at p-values actually does a lot of harm [32]. In some fields, where historically researchers were trained in graphing their data and exploring the actual numerical values, means, and confidence intervals, this practice seems to be decreasing due to the fixation on *p*-values [32]. In computing fields it is not clear that effect size reporting was ever common; Dunlop and Baillie have identified lack of effect size reporting in HCI as "dangerous" [11, p.3] and in the related field of software engineering experiments, a review of 92 experiments published between 1993 and 2002 shows that only 29% of the papers reported estimates of effects[18].

The way forward

To overcome the fixation on *p*-values instead of estimates of effects researchers should report their actual findings, and interpret the numerical estimates of their models or tests. An effect size is "any statistic that quantifies the degree to which sample results diverge from the expectations specified in the null hypothesis" [29, p.991]. Effect sizes have three main uses [29]: Firstly, a prediction of effect size is necessary when planning studies in conjunction with power, sample size and significance criteria (as discussed above). Secondly, it enables researchers to interpret the practical significance of their results because it estimates the magnitude of an effect. Thirdly, reports of standard measures of effect size enable researchers to compare the results from different studies and put their findings in the context of previous work in the literature.

The APA recommend that standard measures of effect sizes should be reported along with *p*-values [1]; they give complementary information. In a study examining the reporting of 855 *t*-tests published in the psychology literature it was found that effect sizes and *p*-values were generally consistent, with large effect sizes corresponding to low p-values. The consistency can in large part be explained by the relatively standardized sample sizes adapted by the field. However, in a small number of cases there were gross inconsistencies between the p-values and standardized effect sizes. These mainly occurred in studies with small sample sizes where the *p*-value was close to .05.

We make a distinction here between standardized measures of effect size – like Cohens's D, eta squared, or the easily interpretable Common Language Effect Size [15] – and non-standardized measures of effect size. The latter are dependent upon the scales by which variables are measured. This latter property makes non-standardized measures of effect size less suitable for comparisons across experiments. However, only non-standardized measures of effect size – *estimates of actual differences in means or parameter estimates in regression models* – can be used to assess the theoretical and practical importance of the quantitative findings of a study.

We believe that standardized effect size measurements can be useful in comparing results across studies. Cohen for example has published useful heuristics for interpreting effect sizes as small, medium or large [6]. *However, it would be of limited value if researchers replaced canned reporting of p-values with canned reporting of Cohen's d or other such statistics.* While standardized effect size measures overcome the confusion of importance and sample size as common for *p*-values, standardized effect sizes cannot, by themselves, be the only outcome of a quantitative experiment. The important point is to consider what the estimations of effect(s) mean in the context of previous work and what the practical and theoretical implications of an effect of that magnitude would be for users or designers.

Example continued: What is really important for usability ratings?

To highlight the importance of an inspection of parameter estimates (e.g., the mean difference or the β's of a model) versus the (often erroneous) interpretation of *p*-values we present a set of hypothetical results obtained from a study similar to that used in the previous examples.

Suppose again the ratings of the usability of two different systems are compared. However, this time we do not only compare Windows Vista en Apple Mac OS-X but we obtain ratings both by novices and by expert evaluators. Table 2 presents the results for this new experiment. The table presents the usability ratings (*N=200*) comparing Vista and OS-X (see also Table 2) when these ratings are provided by novices and by expert evaluators.

	Critiques				
	M. Diff	**SD**	**D**	**T**	**P**
Vista vs. OS-X	*0.91*	*1.5*	*0.6*	*4.29*	*<.001***
Expert vs. Novice	*1.30*	*4.1*	*0.3*	*2.24*	*<.05**

Table 2. Hypothetical results obtained for an experiment. Presented are the mean difference, the pooled SD, Cohen's D, the t-value, and the p-value.

Table 2 presents the (hypothetical) mean-difference between the usability ratings of Vista and OS-X users, and

the mean difference in ratings provided by expert users and novice users. Given the mean differences presented here, the pooled standard deviations, and the equal sample size, both of the effects – that of type operating system and type of evaluator – are statistically significant. Thus, according to most researchers they are both important findings[2].

However, for both theoretical as well as practical purposes it is feasible to evaluate the sizes of the effects of both the operating system as well as the user expertise on the usability ratings that are provided. We have – hopefully – already convinced readers of the inadequacy of *p*-values to make these kinds of judgments. Thus, researchers should *not* decide, based on the lower *p*-value of the operating system factor that this is the most important variable in eliciting critiques.

The third column of Table 2 presents Cohen's d for the two comparisons presented here. Cohen's d is given by the ratio of the mean difference and the pooled standard deviation. This makes the computation similar to that of the *t*-value with the only difference being the exclusion of *N*, the number of subjects, in the equation. Given equal sample sizes in these two evaluations there is a direct relation between the t-value (column 4) and the value of Cohen's d.

Now, should we decide that the operating system is the most important factor influencing people's usability ratings of their systems? The value of Cohen's *d* is higher indicating a higher effect size than for the expertise. However, we think that researchers, knowledgeable of the origin of Cohen's d, should look a bit further. The actual mean difference 'caused' by the expertise of the user as opposed to the type of operating system is far larger. The difference between an expert user and a novice user is around 1.3 points on the 7 point scale, while that for the different operating systems is only 0.91 points. However, the standard deviations indicate that the measures obtained for the different operating systems are more 'consistent' – less spread out – than those obtained for the different user expertise levels.

It is up to the researcher to determine and motivate the conclusions drawn from a dataset like this. However, in this scenario a large standard deviation for expertise is very plausible: actual expertise levels are not binary and thus there is heterogeneity within the expert and novice groups. This argument does not hold for the type of operating system, hence its smaller standard deviation. The actual mean difference however shows that – if assessed accurately – user expertise *could* potentially be a more important determinant of the usability ratings of a system. We believe that discussions like these inform and progress science, rather than the limited interpretation of a single statistic.

We do not mean to imply that all HCI researchers neglect to discuss the size of their effects. For instance in the domain relating to our worked example, a highly cited paper which considers the evaluator effect is very much concerned with the interpretation of effect magnitudes [17]. Within HCI more generally, a good example of a focus on quantitative estimates (rather than just sizeless *p*-values) can be found in the Fitt's law literature. Fitt's law describes the *quantitative* speed accuracy trade-off associated with pointing. The importance of quantitative evaluations when building a science is clear from the status of Fitt's law within HCI research: It presents the only paradigm which consistently fills up at least one session at CHI, and the results are replicated and extended upon frequently.

CONCLUSIONS

Presenting only *p*-values can lead to misleading results with unfortunate real life consequences [9]. The *p*-value often does not inform us about what we want to know, which is generally the probability of the hypothesis given the data. Also, high *p*-values do not imply that the null is indeed true if the power is inadequate, and finally, sizes of effects should be more important than their associated sampling error.

We conclude with a more general criticism of the way theories are developed within HCI. A hallmark of a good theory is that it is highly falsifiable. It should make definite claims about the world, because the more claims it makes, the more opportunities there are to falsify it. A major criticism of the traditional approach to statistics is that it encourages weak theorizing by proposing hypotheses which make vaguely specified claims about the world [9]. In specifying a null hypothesis, the researcher generally predicts no difference between conditions. If this is rejected, the alternative hypothesis is accepted. But the alternative hypothesis that matches this null hypothesis (that there is *some* difference) is vague and underspecified. It rules out only one point where the means are *exactly* the same across conditions. *Any* other relationship between the variables could be true. Seen in this light, *the null hypothesis is intuitively almost always false*, and so rejecting it isn't very informative. A theory which predicts in advance the magnitude of an effect is more useful, and the consideration of estimated effects from the current study in the light of previous findings enables the researcher to contribute coherently to the existing body of work in a field.

Dunlop and Baillie [11] have argued that HCI does not generally attempt replication of previous work, a point which is confirmed in Bargas-Avila and Hornbæk's recent analysis of UX studies [2]. Yet, single studies cannot be taken as the basis for believing a scientific result to be true. It is the pooling of evidence from many studies, often in the form of meta-analysis, that should give researchers

[2] Normally one would analyze this factorial experiment using a method by which dependencies between prototype fidelity and user expertise are also included (e.g. ANOVA). However, we choose to present seperate t-test for ease of understanding of the argument.

confidence in a theory [13] .We should therefore consider: do we as a community want to develop theory through empirical studies (at least as one of the methods in our toolkit)? If we do not, then what purpose is served by conducting traditional statistical tests? If we do, we are more likely to achieve our aims by adopting best practices for the planning, analysis and reporting of empirical studies. Based on the convergence of advice from related disciplines, we offer the following initial recommendations for best practice. We hope that future authors will add to these recommendations.

1. A more specific hypothesis yields more information when it is falsified than a vaguely specified hypothesis. For this reason, *bolder predictions predicting the direction and magnitude of effects would be beneficial* rather than choosing the "safe" null that there is no difference between conditions. For example, a researcher might hypothesize that a shopping website optimized for screen reading software would decrease the average time taken to buy an item by a visually impaired user by one minute over the original version of the website. Such an hypothesis can be evaluated quantitatively, after which qualitative judgments about the importance of an effect this size can be discussed.
2. When planning an experiment, it is helpful to *predict the size of the effect likely to be found*, based on previous findings from related studies if possible. In the above example about the interface for visually impaired users, the researcher could have calculated effect sizes from the descriptive statistics published in previous similar studies, or predicted them from theory or even estimated them from pilot tests in the lab.
3. Deciding on power, significance criterion (alpha value), and effect size in advance enable the researcher to *calculate the number of participants* they require to detect an effect of practical or theoretical importance.
4. If there are practical difficulties in recruiting enough participants, *research teams could consider collaborating for multi-site experiments*. Power can also be increased by careful choice of valid and appropriate measurement instruments.
5. It can be beneficial to *use Bayesian analysis to calculate the probability of the hypothesis given the data* instead of traditional significance testing. This analysis method enables researchers to build on the body of knowledge in the field by incorporating previous results as prior probabilities.
6. We encourage *researchers, reviewers, programme chairs and journal editors to work towards raising the standard of reporting statistical results* in order that future researchers can use this information to inform their own hypothesis generation, effect size estimates and prior probabilities in Bayesian analysis. The guidelines in the 6th edition of the APA publication manual [1] are helpful in this regard. At the very least, the mean and standard error should be reported to enable future researchers to calculate standardized effect sizes. It would be useful for researchers without a strong statistical background if submission instructions for authors included clear guidance to help them enhance their analyses and conclusions.
7. And, last but not least, it is good practice to *interpret the non-standardized sizes of the estimated effects*. If the predicted effect size was found in the example of the shopping web site for visually impaired users, what would this mean? What practical difference would it make to the user experience for members of this target user group? Would a time saving of one minute per transaction be worth the effort it would take to learn how to use the new layout? Some questions of this sort are arguably best answered in consultation with users, emphasizing the need for triangulation between qualitative and quantitative data.

These changes to the best practice within a field will require effort, and may take many years to come to fruition. But if we, as a community, value the tools offered to us by statistical methods, we should do our best to avoid known methodological flaws, and embrace the best practices which are emerging from our sister disciplines. The benefits to HCI will be great in terms of generating a more coherent body of work thus enabling the field to advance more rapidly.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers, Panos Markopoulos, Oliver Lemon, Verena Rieser, Dean Eckles, and Andrew Macvean for their comments on a draft of this paper. Please note that both of the authors contributed equally to this work.

REFERENCES

1. American Psychological Association. (2009). *Publication Manual of the American Psychological Association, Sixth Edition* (6th ed., p. 272). APA.

2. Bargas-Avila, J., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (CHI '11). ACM, New York, NY, USA, 2689-2698.

3. Bayes, T., & Price. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. Philosophical Transactions of the Royal Society of London, 53, 370-418.

4. Cairns, P. (2007). HCI... not as it should be: inferential statistics in HCI research. *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1* (pp. 195–201). British Computer Society.

5. Champely, S. (2009). PWR package. Retrieved from http://cran.r-project.org/web/packages/pwr/index.html

6. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Laurence Erlbaum Associates.
7. Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155. American Psychological Association.

8. Cohen, J. (1994). The Earth is Round (p < .05). *American Psychologist*, *49*(12), 997-1003.

9. Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference* (1st ed., p. 150). Palgrave Macmillan.
10. Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, *6*(3), 274-290.
11. Dunlop, M. & Baillie, M. (2009). Paper Rejected (p<; 0.05): An Introduction to the Debate on Appropriateness of Null-Hypothesis Testing. *International Journal of Mobile Human Computer Interaction*, *1*(3), 1-8.
12. Fowler, F. (1995). *Improving Survey Questions*. Thousand Oaks, California: Sage Publications.
13. Goldacre, B. (2008). *Bad Science* (p. 338). London: Harper Collins.
14. Gray, W., & Salzman, M. (1998). Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Human-Computer Interaction*, *13*(3), 203-261.
15. Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, *79*(2), 314-316.
16. Hansen, W. B., & Collins, L. M. (1994). Seven ways to increase power without increasing N. In L. M. C. (Eds.) & L. A. Seitz (Eds.), *Advances in data analysis for prevention intervention research (NIDA Research Monograph 142, NIH Publication No. 94-3599)* (Vol. 191, pp. 184–195). Rockville, MD:: National Institutes of Health.

17. Hornbæk, K., & Frøkjær, E. (2008). A Study of the Evaluator Effect in Usability Testing. *Human-Computer Interaction*, *23*(3), 251-277.

18. Kampenes, V., Dyba, T., Hannay, J., & Sjoberg, D. (2007). A systematic review of effect size in software engineering experiments. *Information and Software Technology*, *49*(11-12), 1073-1086.

19. Kaptein, M. C., Nass, C., & Markopoulos, P. (2010). Powerful and consistent analysis of likert-type rating scales. *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10* (p. 2391). New York, New York, USA: ACM Press.

20. Katki, H. a. (2008). Invited Commentary: Evidence-based Evaluation of p Values and Bayes Factors. *American Journal of Epidemiology*, *168*(4), 384-388.
21. Kruschke, J. K. (2011). Introduction to Special Section on Bayesian Data Analysis. Perspectives on Psychological Science, 6(3), 272-273.
22. Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in cognitive sciences*, *14*(7), 293-300. Elsevier Ltd.
23. Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological methods*, *9*(2), 147-63.
24. Morey, R., & Rouder, A. (2010). Package "BayesFactorPCL." Retrieved from https://r-forge.r-project.org/projects/bayesfactorpcl/
25. Oakes, M. (1986). *Statistical inference: a commentary for the social and behavioural sciences*. Wiley.
26. Rouder, Jeffrey N, Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225-37.
27. SIGCHI. (2011). More about SIGCHI. *http://www.sigchi.org/*. Retrieved July 8, 2011, from http://www.sigchi.org/about/more
28. Sharp, H., Preece, J., & Rogers, Y. (2007). *Interaction Design* (2nd ed.). Chichester: John Wiley and Sons, Ltd.
29. Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*(4), 989-1004.
30. Wagenmakers, Eric-Jan, Wetzels, Ruud, Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *Journal of personality and social psychology*, *100*(3), 426-32.
31. Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t-Tests. *Perspectives on Psychological Science*, *6*(3), 291-298.
32. Ziliak, S., & McCloskey, D. (2008). *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice and Lives*. (p. 320). Ann Arbor, MI: University of Michigan Press.