HCI... Not As It Should Be: Inferential Statistics in HCI Research

Paul Cairns
University of York
Heslington
York, YO10 5DD
+44 1904 434336

p.cairns@cs.york.ac.uk

ABSTRACT

This paper surveys the use of inferential statistics over the last two BCS HCI conferences and the last year (2006) of two leading HCI journals. Of the 80 papers covered, 41 used some form of inferential statistics. However, all but one had some form of problem of reporting or analysis that undermined the value or the validity of the statistical testing and hence the research findings. This paper discusses the implications of such widespread issues for HCI research and considers approaches for improving the use of statistics in HCI.

Categories and Subject Descriptors

A.1 **[Introductory and Survey].**

General Terms

Measurement, Reliability, Human Factors.

Keywords

Inferential statistics, reporting statistics, assumptions, over-testing, HCI, research methods.

1. HCI AND STATISTICS

Statistics play an important role in usability, as in other human-centred sciences, because they allow researchers to discern patterns in numerical usability data despite the natural variations that we expect to see between people. However, personal experience of the author, both in attending talks and refereeing papers, suggested that perhaps statistical methods were not generally being applied in HCI to the high standard that they should be. This would not be a new insight: Gray and Salzman [10] provide a strong case for the poor standard of research methods, including statistics, in research into evaluation methods. That paper, though, appeared nearly ten years ago, and it may be that attitudes and the behaviour of researchers have changed since then. The goal, then, of the current work is to present a more recent survey of the use of statistical methods, specifically inferential statistics (being the usual understanding of statistics as tests producing *p* values and significance results), and in doing so provide a clear picture of the quantity and quality of statistical methods as used in HCI research. The survey shows that statistics are widely used as an important method in the HCI research methods canon. Sadly, though, the use of statistical analysis is greatly undermined by poor reporting, inappropriate analysis or insufficient analysis.

This paper outlines the key findings of the survey and suggests recommendations for the improvement of statistical methods as used in HCI.

2. APPROACH TO THIS STUDY

To evaluate the quality of a statistical analysis, it is necessary to have some standard against which to judge the analysis. The approach taken here is in the tradition of psychology, however the standard in psychology is usually far more implicit than explicit. Excellent introductory textbooks such as Rowntree [19], or more advanced ones such as Greene and D'Oliveira [11], do tell students how to perform statistics in a particular manner, but the full reasons for such procedures are not always explained. Psychology students tend to acquire the good practice by being exposed to it in their studies and corrected if they do not do it 'right' without ever being told what 'right' is.

To be explicit for this paper, the standard used here is based on null hypothesis significance testing (NHST), as opposed to other methods such as confidence measures and Bayesian approaches. NHST is the most commonly encountered form of statistical inference and is what is usually associated with producing a null hypothesis, then testing it to give some statistic such as a *t* value, and then turning the statistic into a *p* value.

The standard for reporting NHST in psychology is specified in the APA manual [2], and indeed this seems to be a standard also used across other disciplines in education, health and the social sciences.

An additional feature of standard NHST is that NHST is not considered as proof of an alternative hypothesis, but is an argument form that lends evidence towards the alternative hypothesis [1]. The argument of NHST relies on the prediction of an unlikely outcome that may happen by chance but that the researcher predicts is due to the experimental manipulation. Every effort is made to ensure a chance outcome except for the particular manipulation. The data are gathered and tested to see how likely they are to have occurred by chance. When the data are unlikely to have occurred by chance, usually with a probability of less than 0.05, they are said to be significant and this gives weight to the researcher's theories and insights. To illustrate this argument more concretely: throwing a double six with two dice is unlikely; the probability is 1 in 36, which is less than 0.05! But no-one is surprised when double six occurs in dice games as it is just what happens now and again. However, if a person takes a pair of dice and predicts that the

Published by the British Computer Society
People and Computers XXI – HCI… but not as we know it: Proceedings of HCI 2007
Linden J. Ball, M. Angela Sasse, Corina Sas, Thomas C. Ormerod, Alan Dix, Peter Bagnall, and Tom McEwan (Editors)

next throw will be a double six and they are right then anyone watching is surprised. In fact, they would be wise to suspect that the thrower is up to something simply on the basis of that *one* throw. In the same way, the unlikely outcome of an experiment should make people think that the researcher knows something about what is going on in the experimental set up, that is, they have a good theory of how things work (provided the experiment has been done well).

This argument form has implications for the conduct of statistical analysis; for an analysis to be sound, it is necessary that in the tests performed the probabilities of outcomes are accurately reflected in the *p* values produced by the tests. If this is not the case, then the NHST argument form is severely weakened. This will be discussed in more detail in the context of the survey findings.

3. PAPERS CONSIDERED

As the intended outlet for this paper is the BCS HCI, the survey considered the two most recent years of BCS HCI conferences, namely 2005 and 2006 [17, 5]. Only papers appearing in volume 1 of the proceedings were considered as these are intended to be reports of mature research that meet a higher standard than that covered in posters and short papers. It should be noted that in 2006, the BCS HCI included six short papers as these were of a particularly high standard and merited inclusion in volume 1. Two conferences were used in case the standard was affected by external influences specific to one particular year, say perhaps because of location of the conference or clashes in timing with a related conference. Focusing only on this conference series may still suggest that any issues identified in the survey were only particular to the conference series. Thus, two major journals, Human Computer Interaction (HCIJ) and ACM Transactions on Computer Human Interaction (TOCHI), were taken as comparable outlets of high level research in HCI, but that were independent of this or any other conference. The papers appearing in all issues of both of these journals for the year 2006 were considered.

The purpose of this survey is not to make direct, public criticism of individual researchers. In what follows, individual papers are not singled out or directly referenced but are only discussed in aggregate. Hence, they are also not listed in the references at the end. If the reader would like to know more details of the individual papers or the analysis made, they are welcome to contact the author.

The two proceedings and two journals were made up of 80 research papers of which 41 used inferential statistics in some form or other, see Table 1 for the breakdown by outlet. Thus around half of all the papers considered use some form of statistical tests, and this is roughly the same for all four outlets considered.

Table 1. Number of papers considered by outlet.

Outlet	All papers	Papers using statistics
HCI06	20	13
HCI05	31	13
HCIJ (2006)	10	5
TOCHI (2006)	19	10
Total	80	41

Table 2. Number of papers using at least one instance of a particular test.

Outlet	F	t	Multivariate	Non-parametric	Other
HCI06	9	3	1	5	5
HCI05	3	9	0	3	3
HCIJ	2	2	2	2	2
TOCHI	6	6	2	2	2
Total	20	20	5	12	12

As might be expected, the most popular tests were *F* tests (ANOVA) of various designs and *t* tests (some implemented as *F* tests) including using both tests as follow-ups to omnibus tests. There were also a smaller number of other parametric, non-parametric and multivariate tests. The number of papers using a particular test is given in Table 2. It should be noted that most papers used more than one instance of any particular test and frequently used more than one type of test.

There were additional papers that did use numerical data from questionnaires and studies but used only summary statistics such as means and percentages. Arguably, these papers could have benefited from using inferential statistics, but as the purpose of this survey was to describe what was being done rather than what was not being done, these papers were not considered further.

4. SURVEY FINDINGS

Across the papers considered, common problems occurred in the analysis of data using statistics. These can be broken down into problems of:

1. Reporting
2. Checking assumptions
3. Over-testing
4. Using inappropriate tests

These will be considered in turn in the following sections. Out of the 41 papers using statistics, only one paper did not cause concern over any of these features. Unfortunately, this was not due to particularly careful execution of statistics but because the paper deliberately took an exploratory approach to analysing a questionnaire and appropriately felt that formal statistics were unnecessary except for two Spearman rank-order correlations to show relationships between aspects of the questionnaire.

It is also fair to mention that three papers contained a substantial amount of good statistical analysis, but as they were from the HCI conference, they were limited in page length. This meant that some aspects of the analysis were not covered in detail and therefore failed to report or discuss the tests in sufficient detail to understand what had been done and why. In particular, two of these papers were intended as short papers and so were even more restricted which may have led to their reporting problems. Even so, the result was a lower standard of statistical analysis, albeit amongst some generally good analysis, and as such these papers will be included in the findings below.

4.1 Reporting

One of the dominant problems of the papers was their reporting of the statistical analysis. The APA Manual [2] guides psychology researchers to “include sufficient information to allow the reader to fully understand the analyses conducted and possible alternative explanations for the results of these analyses”. The manual also acknowledges that different statistical tests require different sorts of reporting.

Twenty five of the papers using statistics failed to meet the basic level of reporting required by the APA. Three papers clearly did do statistical tests as they reported *p* values, but as there was no other information it was impossible to know what tests had been done, what the actual statistics were, and, indeed, what the values were that had been compared with the statistics. In four further papers, it was clear that particular tests had been done and the *p* value was given but there was no further information. Three of these are the conference papers mentioned earlier and hence the lack of adequate reporting could have been an editing decision. In these cases, the fact that many other tests had been reported well gives confidence in the findings of these under-reported tests but does not allow the readers to make up their own minds. The fourth such paper appeared in HCIJ, so had no such restrictions, nor did it have other statistical tests done well that would provide confidence in the poorly reported ones.

Generally though, most papers did report the test that was used together with appropriate statistical values. However, another problem was that of reporting enough to enable readers to assess the analysis for themselves. Key to this is having the actual values being compared. In particular, for a *t* test or ANOVA, these are the means of the groups that are being compared and the associated standard deviations of each group. Four papers failed to report the means for *t* tests or ANOVAs, though one paper did report them graphically. Fifteen papers failed to report the standard deviations. Oddly, two papers reported the standard deviations in some tests and not in others. Some papers also neglected to make any report of tests that were not found to be significant, though they mentioned that such tests had been done. It was hard to give a clear figure to this because obviously some authors may have done tests and entirely omitted mention of them in the paper. However, the presence of these few such papers does suggest that it is generally viewed as unnecessary to report non-significant results. However, on the basis that the reader should be able to decide for themselves, these tests should be reported just the same as those that came out significant.

4.2 Checking Assumptions

All statistical tests make basic assumptions on the nature of the data being examined. At the very simplest level, the data type (categorical, ordinal etc.) is considered, but there may also be distributional assumptions that are important for a test to give an accurate *p* value. For NHST to be valuable as an argument, every effort should be made to ensure that the *p* value produced genuinely reflects the probability of achieving the data obtained by chance. Violation of the assumptions of any statistical test can produce *p* values that bear little relation to the actual probabilities of outcomes, and hence comparison to the significance level of 0.05 is meaningless.

For *t* tests and ANOVA, the assumption is that within each group considered, the data are normally distributed. Moreover, the variance (being the square of the standard deviation) of the data in each group is the same; this is the ‘homogeneity of variance’ condition. It is generally stated in statistical textbooks that *t* tests and ANOVA are robust to deviations from these assumptions, provided the group sizes are approximately equal. For *t* tests this is indeed the case [20], but for ANOVA it is not [7]. The robustness statement is grounded in earlier work [4], but the variations in variance used there were actually quite small in relation to actual variations observed in practice. Even modest ratios of largest variance to smallest variance of 4:1 are enough to make the significance level more liberal than the stated 0.05. Thus it is important to make checks on both normality and homogeneity of variance to ensure that the conclusions from an ANOVA are sound.

Nine papers performed ANOVA tests but made no consideration of the assumptions underlying the test. Most of these also failed to report standard deviations so that it is not possible for the reader to check that the assumptions are plausible. Unfortunately, stating the standard deviations reveals the severity of the problem. One paper had an ANOVA on data where there was a ratio of 50:1 between the highest standard deviation and the lowest standard deviations. This corresponds to a ratio of variances of 2,500:1! Moreover from the stated figures, it was clear that the group means correlated with the group standard deviations, which is also a feature of non-normal data (most likely an exponential distribution). It can safely be asserted that the conclusions drawn from that particularly ANOVA are severely in doubt.

Five papers used multivariate statistical tests such as MANOVA, multiple regression and ANCOVA. These tests are highly dependent on their distributional assumptions, including issues of multivariate normality, correlations between the measures and homogeneity of covariance (the multivariate form of homogeneity of variance in ANOVA) [12], and thus it is essential when using such tests to check that the data meet the assumptions [15]. None of the five papers reported any such analysis or gave indication that any such analysis had been done.

4.3 Over-Testing

It is well recognised that it is important not to over-test data, but it is not always clear why, after all, the data will not change just because more tests are done on them. However, from the NHST argument form, repeated, unplanned testing of the data is like rolling two dice many times – the more the dice are rolled, the less surprised anyone should be when a double six comes up. Similarly, if the significance level is set at 0.05, then this is the probability of the data occurring by chance when there is no experimental effect, namely one in twenty times. The more tests that are done on a particular dataset, the more likely it is that some chance variation will be extreme enough to seem like significance.

Overall. papers did not report multiple tests on different configurations of the same data. Only one paper made multiple comparisons in one particular analysis of several different groups in the same dataset. No adjustment was made for these multiple tests. Thus it seems researchers are cautious on the whole about over-testing. However, this is not the only way in which over-testing can happen.

One subtle problem is the use of ANOVA. In general, ANOVAs are used to avoid over-testing by doing an omnibus test of the differences between all groups rather than many pairwise comparisons between the groups. In a one-way ANOVA this is indeed the case. However, if the dimensionality of the ANOVA goes up, so too does the number of tests actually performed. A two-way ANOVA produces three *F* values, a three-way ANOVA produces seven and a four-way produces fifteen. These are undoubtedly fewer tests than all the possible pairwise comparisons between all the groups, but for three-way and four-way ANOVAs there are still enough comparisons to risk over-testing the data. The probability of finding at least one significant *F* value by chance in a four-way ANOVA is 0.54, and for a three-way ANOVA is 0.30. These are both considerably higher than the ostensible significance level of 0.05, and hence undermine the evidence provided by the significance of any one particular *F* value. Three papers performed four-way ANOVAs, and a further four performed three-way ANOVAs, all without consideration of over-testing.

Another issue is when a dataset consists of multiple measures made on the same individual, for example, a person performs a task and the study measures time, accuracy and correctness of performance (This is actually a common design corresponding to the usability criteria of efficiency and effectiveness). It seems reasonable to perform tests on each of these measures separately, but there may be relationships between the data that equate with actually testing slightly different forms of the same data. In this situation, this is over-testing, and chance variation between the measures may push a test into significance.

Twelve papers did exactly this sort of testing, with many different tests performed on many different measures without consideration of the relationship between them. Of course, it may be that there is no relationship between the measures, but this could have been checked using correlations.

4.4 Using Inappropriate Tests

On the whole the papers did tend to use appropriate tests to evaluate the data, but there were definitely situations in which the choice of tests used was either unclear or could be called into question.

Two papers clearly stated using specific tests to make particular comparisons, but the choice of test was questionable. These papers were also the conference papers that had limitations on space, and hence it may be that the details of these tests were lost in achieving the page limit.

Another three papers made multiple pairwise comparisons between groups where an ANOVA would have been more usual and more appropriate to avoid the risk of over-testing. In all three papers, though, the number of comparisons that were made were few, and hence the advantage of an ANOVA may be minimal, particularly if pairwise follow-up tests were required to interpret the results [23].

Two papers mixed parametric and non-parametric tests on the same dataset. This is quite strange because if the data really were non-parametric then parametric tests should not have been used at all. At the very least, some justification of the change of style of test might have been useful so that the reader could be clear on what motivated the decision to use different types of test.

Whilst these examples of inappropriate testing could be equivocal, especially if the reporting were better, when it came to three modelling papers, the use of statistics was clearly incorrect. The goal of these three papers was to compare the performance of users against the predictions made by models. In two of the papers, the user data were expected to fit a trend as predicted by a model. To evaluate the trends, an ANOVA was used to compare the values, and a significant result was interpreted to mean that the trend was correct. This is not what an ANOVA can show. The significant result of an ANOVA means that at least one of the group means is different from the others at a level that is unlikely to be chance. This says nothing about the direction of the difference, nor the overall pattern of differences of all of the groups. What is needed to compare a group of data points to a predicted trend is a contrast [8]. This is a parametric test comparable to an ANOVA, but where the researcher makes an *a priori* prediction about the relative distributions of the group means. A significant result from a contrast would allow researchers to draw conclusions about trends seen in the data.

In the third modelling paper, *t* tests were used to show that actual values obtained from users were *not* significantly different from the predictions of a model. This seems acceptable, and such tests do add weight that the model is making correct predictions because significant differences would highlight problems, but actually the use of *t* tests in this way is inappropriate. The assumption of such a *t* test is that the obtained mean is already from a population which has the predicted mean. A significant result means that a sample *does not* come from a particular population. It says nothing about the likelihood of the sample actually coming from the population without some due consideration being made about the size of the effect and the power of the test. This is exactly analagous to the difference between a type I and type II error [19]. Admittedly, there is no immediately obvious statistical test that could be used to evaluate this. A correlation or a contrast could be used to match the trend of values to actual values, but it would not say how close the actual values were to the predicted values. It is also possible to do null testing but this is rarely done and even more rarely found in textbooks.

4.5 Summary

Disappointingly, the 41 papers reviewed here bore out the initial suspicion: within HCI the standard of statistical analysis is generally quite low for providing convincing results based on NHST. Most of the problems relate to adequate reporting, but there also seems to be a general lack of concern for the validity of statistical tests in terms of the assumptions required for validity, the number of tests that ought to be performed and even the choice of tests. This low standard did not seem to be particularly dependent on the choice of outlet (though three papers, as mentioned, may have suffered from being in a conference with page limits rather than in a journal) as can been seen in the breakdown in Table 3.

Table 3. Number of papers with at least one instance of a particular statistical problem.

Outlet	Report	Assumptions	Over-testing	Inapt test
HCI06	9	5	5	4
HCI05	7	2	4	5
HCIJ	2	3	1	1
TOCHI	7	6	5	2
Total	25	16	15	12

5. IMPLICATIONS

Given that the goal of these research papers is to further research in HCI, the first question must be: what is the impact of the quality of the statistical analysis on the research findings? This is a matter of perspective.

Only two of the 41 papers considered could be said to have broader social impact, as they consider issues of accessibility and of crime-scene investigation. Whilst they did have errors, the errors were errors of reporting, and it may well be the case that the findings reported are perfectly sound. Thus, from this perspective, the poor quality of statistical analysis across the 41 papers is not hugely problematic.

Another perspective might be to assess the impact in terms of those papers that are highly cited and hence it is more important that they are correct. However, both these views denigrate the fact that all of the papers were intending to make a contribution to HCI, and just because they were not so socially oriented or widely cited is not to say that the work was any less good or important. The perspective required here, then, is to consider the impact of the statistical analysis in terms of contribution to knowledge of HCI, regardless of the purpose of seeking that knowledge. Science has set itself up so that there are methods and tools that if used correctly are accepted by the research community as leading to sound knowledge [16]. This is a relativist stance and, as such, means that the tools and methods are open to negotiation and development and with them so too is science.

From a within-paradigm perspective, statistical methods only produce sound results within certain parameters. The NHST makes clear what those parameters are, and it is the conventional framework for using statistical methods (though this is rarely made explicit). If HCI research is functioning within the NHST paradigm then the standard of statistical reporting and analysis is poor. A large majority of papers have substantial problems, which means that their contribution to knowledge must be strongly called into doubt. HCI researchers should not feel alone in this. It is well recognised within psychology and social sciences that there is a poor understanding of significance testing amongst students, teachers and researchers [9, 13]. However, in those disciplines, the ritual application of correct methods such as correct reporting and checking of assumptions does, to some extent, protect researchers and their readers from the misinterpretation of experimental data.

It could be, though, that HCI does not regard the NHST as the most useful framework for statistical analysis in HCI. In which case, what framework is? None of the papers even suggested that an alternative framework for statistical argument was being considered. The only paper that declined to use NHST was the only one not having any statistical problems! To be fair, none of the papers explicitly stated that they were using the NHST framework either. However, the approaches used and conventions observed did suggest that the studies were conforming to the usual style of psychology experiments, and hence to the implicit NHST framework. This may not be a fully reliable inference since, as is common practice in statistical analysis, no statistics textbooks or papers were referenced to support the analysis done. Thus, the papers surveyed were not calling into question the NHST framework but nor were they meeting the standard that it sets.

This does not mean that HCI as a community accepts NHST. Indeed, the psychology research community has been strongly questioning the value of NHST in psychology for some years now [6], and calling for a more meaningful reporting of statistical inference based on effect sizes, confidence intervals and Bayesian reasoning [9]. However, the starting point for this is the correct execution of statistical analysis within the NHST framework, then extending beyond that to provide conclusions that are more useful to subsequent researchers. There seems to be no suggestion within HCI that a more rigorous approach is required since the approaches currently being used are far from rigorous.

There is a still more fundamental question: is statistical analysis appropriate in HCI at all? It could be that with more recent emphasis in HCI on user experience, for example [3], it is not possible to consider the impact of designs on the "mean experience". Instead, HCI should be moving away from quantitative approaches to a much more qualitative approach that acknowledges the individual's relationship with technology rather than some aggregate, average measure of user interaction. However, the evidence from this survey is that statistics is considered necessary since around 50% of all the papers made some use of statistics. Also, although there is an emphasis on experience, this is not to say that all HCI researchers are abandoning more traditional work-based or task-oriented research, and for them, some form of quantitative analysis is going to be useful.

A broader implication of this survey is that regardless of why the statistical analysis in the actual papers is weak, it is not solely the fault of the authors. In all four outlets, the reviewing process is substantial, involving usually three and possibly more referees, as well as editors, which in the case of the journals are usually specifically assigned to each paper. Thus, not only are authors producing weak statistical analysis but this is being accepted by a large number of their peers. This suggests that the knowledge of how to do appropriate statistical analysis is poor in a much larger portion of the HCI community than just those who have been (indirectly) referenced here.

6. RECOMMENDATIONS

It seems at the very least, that the HCI community, or even just the BCS HCI conference, could set a standard for adequate reporting and execution of statistical analysis. The APA Manual [2] sets a particular standard of reporting that is valuable in psychology (though not with universal agreement), and this could be adopted. Alternatively, some other standard could be devised that better reflects the concerns particular to HCI. Also, referees could be made to make explicit consideration of the statistical analysis and assess its adequacy. If a particular referee does not feel able to do so, then this would be visible to the editors in the review, and they could ensure that at least one referee has made a thorough consideration of the statistics of each paper.

Additionally, in these times of free and easy flow of data, it should be possible for journals and conferences to set up a repository of data associated with each paper, and it is a requirement of publication that such data are submitted. Thus, whilst authors and referees strive to do good statistics to the limit of their knowledge, all readers are ultimately able to check the quality of the statistical analysis by doing it for themselves.

The underlying problem seems to be due to a broad lack of adequate statistical education. The basic knowledge of how to do good statistics is not at all ingrained in HCI researchers in the way that it is in psychology researchers. There are good reasons for this [22]. HCI draws on psychology but it also draws on computer science, design, social sciences, business and so on. Each of these disciplines have their own tools to arrive at secure knowledge, and it would be challenging, to say the least, to teach all of these methods to the same high standard as would be expected of statistics. Thus, many researchers are left to teach themselves statistics, but this is not easy because a lot of the expectations articulated in this paper are actually implicit and arise from the culture of practising statistics rather than being found in books. Specifically, there are no excellent statistical textbooks that reconstruct the full subtlety of statistical analysis and its underlying assumptions whilst remaining accessible. Abelson's book [1] comes very close, but even then there is perhaps the expectation that readers are quite fluent with how statistics are done and the book is to help readers to deeper insights.

Thus, it seems that some HCI specific education could be of great benefit. Conferences offer a good opportunity for tutorials and doctoral programmes where statistical methods can be taught in a manner best suited to HCI. However, particularly with doctoral students, the impact on the practice of HCI research could be slow, effectively waiting for a new 'generation' of HCI researchers. Journals and conferences, though, could offer some sort of training for their editorial and reviewing teams. The goal may not be to make everyone an expert in statistics, but rather to be able first to identify the need for a particular statistical expertise in the reviewing team, and then to be able to find a team member with that expertise. Arguably, such an approach to reviewing could be generally beneficial in accounting for the multidisciplinary nature of HCI research.

More immediately, the survey seems to suggest that some problems could be remedied in any future work. Reporting statistics is obviously straightforward, particularly if the statistics have been done well in the first place, and a guide such as the APA manual [2], or one of the many websites that explain the manual with examples, is a good starting point (for example [14]). There are also research gains to be had from better reporting as well. Whereas under current practices, non-significant results are little better than a footnote of failed attempts, properly reported non-significant results can help future researchers to provide estimates of effect sizes and associated confidence intervals. This can be of relevance in understanding how to design future studies in the given area, and could be combined with significant results to provide tighter estimates on population means than is possible with a significant result on its own [21].

In terms of the actual analysis done by researchers, it is natural to think that HCI is complicated with so many aspects of context influencing interactions. But accounting for this by increasing the complexity of statistical tests is not necessarily the answer. Multi-way ANOVAs lead to a risk of over-testing. This is not to mention trying to interpret and then communicate what a three-way interaction effect really means when there are also two-way interaction effects and main effects present [18]. Also, multivariate methods bring their own hazards, in particular, in terms of their distributional assumptions. Even just measuring more than one thing can introduce the problem of how to analyse the many measures without risking over-testing or using multivariate methods. Thus, whilst it may seem tempting to use such methods and design experiments accordingly, actually they raise more problems than they solve, at least on the evidence from the papers surveyed here. A useful rule of thumb from this is to keep statistical analysis as simple as possible. Of course, this is bearing in mind Einstein's caveat that it should be as simple as possible but no simpler.

In the longer term, it seems that the HCI community should engage in a discussion of what it means to do good statistical analysis in HCI. The NHST only offers one sort of framework, and others may be better or more acceptable depending on the goals for HCI research. Perhaps with this new framework and some of the recommendations such as above, we will produce a new form of statistical analysis. It will be HCI not as we know it but hopefully as we feel it should be.

7. ACKNOWLEDGMENTS

Many thanks to Simon Attfield, Jeremy Gow, Harold Thimbleby and Eduardo Calvillo for some very thoughtful comments on this paper.

8. REFERENCES

[1] Abelson, R. J. *Statistics as Principled Argument.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.

[2] American Psychological Association. *The APA Manual.* APA, 2001.

[3] Blythe, M. A., Overbeeke, K., Monk, A. F., and Wright, P. C. (eds.), *Funology: From Usability to Enjoyment.* Kluwer Academic Publishers, 2004.

[4] Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality variance in one-way classification problems. *Annals of Mathematical Statistics, 25* (1954), 290-302.

[5] Bryan-Kinns, N., Blandford, A., Curzon, P., and Nigay, L. (eds.), *People and Computers XX (Proceedings of HCI 2006).* Springer, London, 2006.

[6] Cohen, J. The earth is round (p<0.05). *American Psychologist, 49,* 12 (1994), 997-1003.

[7] Coombs, W. T, Algina, J., and Oltman, D. O. Univariate and multivariate omnibus hypothesis tests selected to control type I error rates when population variances are not equal. *Review of Educational Research, 66,* 2 (1996), 137-179.

[8] Furr, R. M., and Rosenthal, R. Evaluating theories efficiently: the nuts and bolts of contrast analysis. *Understanding Statistics, 2,* 1 (2003), 45-67.

[9] Gigerenzer, G. Mindless statistics. *Journal of Socio-Economics,* 33 (2004), 587-606.

[10] Gray, W. D., and Salzman, M. C. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13,* 3 (1997), 203-261.

[11] Greene, J., and D'Oliveira, M. *Learning to Use Statistical Tests in Psychology.* Open University Press, Milton Keynes, UK, 1999.

[12] Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. *Multivariate Data Analysis (5th edition).* Prentice-Hall, New Jersey, 1998.

[13] Haller, H., and Krauss, S. Misinterpretation of statistics: A problem students share with their teachers? *Methods of Psychological Research Online, 7,* 1 (2000).

[14] Hesson-McKinnis, M. *Reporting Statistics in APA Style.* Accessed March,2007, http://www.ilstu.edu/~mshesso//apa_stats_format.html

[15] Keselman, H. J., Huberty, C. J, Lix, L.M, Olejnik, S., Cribbie, R.A., et al. Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research, 68,* 3 (1998), 350-386.

[16] Kuhn, T. S. *The Structure of Scientific Revolutions (3rd edition).* University of Chicago Press, Chicago, IL, 1996.

[17] McEwan, T., Gulliksen, J., and Benyon, D. (eds.), *People and Computers XIX (Proceedings of HCI 2005).* Springer, London, 2005.

[18] Rosnow, R. L. and Rosenthal, R. If you're looking at the cell means you're not looking at *only* the interaction (unless all the main effects are zero). *Psychological Bulletin, 110,* 3 (1991), 574-576.

[19] Rowntree, D. *Statistics without Tears.* Penguin Books, London, 2000.

[20] Sawilowsky, S. S., and Blair, R. C. A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin, 111,* 2 (1992), 352-360.

[21] Smithson, M. *Confidence Intervals.* Sage Publications, London, 2003.

[22] Thimbleby, H. Supporting diverse HCI research. In *Proceedings of the 2004 BCS HCI Conference, Volume 2*. Research Press International, Bristol, UK, 2004, 125-128.

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