From Mice to Men – 24 years of Evaluation in CHI

Louise Barkhuus
University of Glasgow
17 Lilybank Gardens
Glasgow, G12 8QQ
United Kingdom
barkhuus@dcs.gla.ac.uk

Jennifer A. Rode
Donald Bren School of Information
444 Computer Science Building
Irvine, CA 92697
USA
jen@ics.uci.edu

CHI 2007, April 28 – May 3, 2007, San Jose, USA
ACM 1-xxxxxxxxxxxxxxxxx.

Abstract
This paper analyzes trends in the approach to evaluation taken by CHI papers in the last 24 years. A set of papers was analyzed according to our schema for classifying type of evaluation. Our analysis traces papers' trend in type and scope of evaluation. Findings include an increase in the proportion of papers that include evaluation, and a decrease in the median number of subjects in quantitative studies. We also critique the types of subjects, in particular an over reliance on students, and lack of appropriately gender balanced samples. We contextualize these findings in historical trends as we move from machines intended for the technical elite in laboratories to computers integrated into the daily life of everyone.

Keywords
Evaluation, Qualitative, Quantitative, History, Gender, User Experience, Meta-HCI.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. K2. History of Computing: Theory.

Introduction
An important part of HCI is evaluation—evaluating new application and technologies, as well as, the

environments in which they are integrated. Since the early days of HCI where human factors papers looked at users' performance on different computers [6, 28], to today where ubiquitous computing applications are being tested in the wild [3, 9], evaluation has stood out as an integral part of our field. HCI is as much about evaluation as it is about development. In fact, sufficient validation is one of the review criteria for CHI publications describing systems of applications [7]. It seems as evaluation has become synonymous with presenting work at CHI, however, this has not always been the case, as we will demonstrate in this paper.

Despite evaluation being a coherent part of CHI, it has not been systematically studied. Few exceptions exist: First, Gray and Salzman [14] critiqued five studies, which compared the effectiveness of different usability evaluation methods. Secondly, Grudin's work provides a valuable historical overview of trends relating to evaluation like who is doing the evaluation and who we consider our users to be [18]. Our work looks at how practitioners and academics have used evaluation techniques through the past 24 years. By surveying these techniques, we hope to explore how evaluation has been used, how quantitative and qualitative evaluation has played together and what type of subjects CHI evaluation has used. Like Grudin we use a historical lens incorporating trends in the field, but also, projecting what these trends mean for members of the CHI community. More specifically we try to answer three questions:

1. How has the role of evaluation changed through the last 24 years of CHI as technology evolved?
2. How has empirical evaluation, the most prominent type, developed in scope?
3. What type of subjects has CHI used through time?

Although it seems as a large task to answer these questions in detail, we aim to present the first steps in the direction of exploring evaluation methods in CHI. To explore the strengths and shortcomings of evaluation in HCI, we have analyzed a set of representative CHI papers for evaluation type and scope. Although CHI is not the only forum for HCI research, we find that this community is particular influential and it is widely acknowledged as the most recognized forum for HCI work [17, 40]. It also provides the most familiar reference point for analysis and presentation in this forum. Before providing an overview of human factors research and its introduction of evaluation with human subjects to the field of computing, we briefly visit CHI's review process, an essential parts of the background material for our survey.

Background

CHI and its Review Process

Like most other research communities, the CHI community uses peer-reviewing to evaluate new research, meaning CHI papers are reviewed by 'insiders', HCI researchers or practitioners themselves. However, the definition of an HCI researcher is not straightforward and the review body has transitioned from being experts selected by the community to a self-selected body including junior members such as students. For instance, nine years ago a discussion of what constitutes an adequate review body occurred, causing SIGCHI to open up the review pool from invited

pre-published researchers to anyone who might sign up [28]. Although the review process was still structured by associate chairs they now had a different pool to choose reviewers from.

Prior Classification of CHI Literature

Our work is not the first to survey CHI publications. In 1990, Wulff and Mahling examined seven consecutive years of CHI conference publications for topical trends [40]. They categorized papers according to contribution. Evaluation was only one of many categories, and all evaluation papers were lumped together. Our work, rather than trying to classify the CHI literature as a whole, focuses instead on examining the nuances and sub-categories within the evaluation literature.

In the next section, we trace broadly the use of evaluation in the CHI community over the last 24 years, and how it has responded to various influences.

Emergence of Computing System Evaluation

Prior to the mid-seventies, evaluation in computer science was mostly concerned with computer performance [5], but as user interfaces became more prominent and important to computer use, resulting in a much more diverse set of users, researchers began to look at the evaluation of user performance. Researchers who focused on this were mainly psychologists who brought in evaluation methods based on quantitative experiments [18]. Early HCI research, like today, was concerned with interaction between users and the computer, however, this included programming and command based tasks. Early HCI research for example evaluated how users learned command lines [19] and processed tacit programming knowledge [36] as well as it studied how users perceived and handled the user interface [20]. The latter became particular important as graphical interfaces emerged around the early eighties. It was not until then that the modern conception of the end-user developed, with the advent of the personal computer [18]. One of the major changes in HCI research is that concepts described are increasingly accompanied with user evaluation. In fact, as we will provide evidence for later in this paper, evaluation in the eighties was often provided solely by describing the system or conceptual model in detail where this is rarely the case with more recent HCI research.

The nineties saw an increasing debate in terms of evaluation of usability, largely due to the advent and availability of the modern Internet. Some advocated discount usability [31] where others called for more thorough controlled usability studies [15]. Around this time papers solely devoted to usability methods themselves therefore increased slightly in numbers.

With the integration of other fields into HCI such as anthropology and sociology, as well as the influence of related research areas such as CSCW, different evaluation methods have recently been used and become subject of debate. Sengers and Gaver, for example, argue that evaluation should be broadened out to allow for multiple interpretations rather than focusing on one single interpretation or task [35]. Their view is that classic usability testing is 'outdated' and insufficient for new domains such as domestic and public environments and that evaluation needs to encompass dynamic feedback and users' broader understanding of the system.

Method

We chose to limit data collection and analysis by selecting a subset of CHI papers rather than analyzing them in total. We chose a sample because we felt that the insight gained from a potential analysis of all CHI papers would not outweigh the massive task of indexing 1569 papers. So, only selected years of research were studied in depth. Wulff and Mahling indexed 360 papers, representing seven consecutive years of CHI papers, whereas we chose a sample that permitted a longitudinal study of all 24 years. We choose five years worth of papers, most with a six-year gaps. They years chosen were 2006 (118 papers, acceptance rate 23%), 2000 (72 papers, acceptance rate 21%), 1994 (70 papers, acceptance rate 27%), 1988 (39 papers, acceptance rate 21%) and 1983 (59 papers, acceptance rate 34%). 1983 was chosen instead of 1982, which was the very first CHI conference (however, this is still debated in the community) because 1983 had an acceptance rate much closer to the future years and therefore seem more representative than 1982.

Taxonomy for Paper Indexing

When analyzing the type of evaluation used by papers we first classified papers as either containing evaluation or not containing evaluation (systems, algorithms and applications without evaluation, systems which were subject to non-rigorous opinion-based evaluation of the 'do you like it' flavor, theory papers, surveys, new usability and design techniques, models of user behavior, or papers on design process).

Papers that contained evaluation were classified along two axes—in regards to whether the study involved users or not (empirical or analytic), and methodological approach (qualitative or quantitative or a combination of both). See figure 1 for more detail. Our classification schema was high level, and we recognize that there are many methodological nuances within both qualitative and quantitative studies. However, the aim of our survey was to trace trends rather than subtle variations in evaluation.

Unlike Wulff and Mahling's classification, we classified papers presenting system design *including* evaluation alongside papers that presented evaluation as their primary contribution (for example studying the use of an existing system or technical environment). Our focus was not on trends in CHI papers as a whole, but a longitudinal study of trends in CHI papers with respect to evaluation, focusing on the nuances there-in and the changing trends.

In addition to indexing papers according to the taxonomy described above, each paper's contribution, as well as its number and type of subjects were described. The authors did the indexing with the assistance of a colleague; to insure consistency, all final indexing was reviewed by the first author. The taxonomy itself was subject to iteration and refinement to accommodate the differences in the evaluation styles historically in CHI. The process was very similar to refining categories for open coding of ethnographic data.

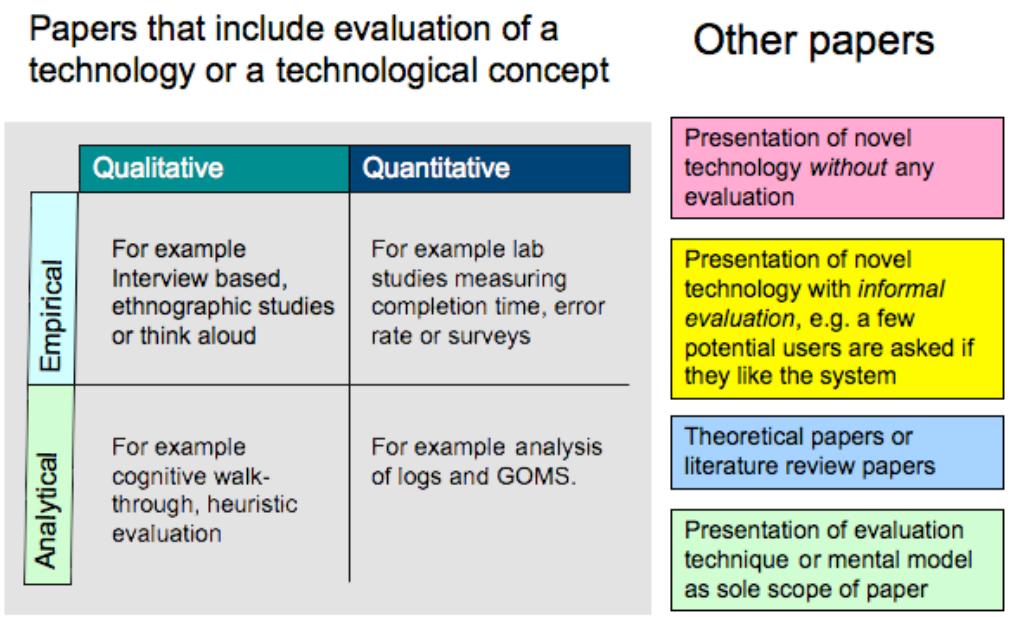

Figure 1: The taxonomy for indexing our data set.

Our analysis resulted in an incredible amount of data, of varying degrees of interest. We now turn to each of our questions, the first looking at the changing role of evaluation throughout the last 24 years.

Changes in the Role of Evaluation

Evaluation is a core part of HCI and has always been present in research, however, only recently has this become an actual criteria for publication. It is telling that Wulff and Mahling where able to classify artifacts and evaluation independently in their study, although they acknowledged that many of their 'artifact building' papers included evaluation, commenting "As more and more papers both report on and advance, and then evaluate it, the pure evaluation category becomes less important and we almost need new categories like 'artifact building plus evaluation'" [40]. Evaluation trends reflect this as we see a transition from a diverse number of informal methods and papers without evaluation in the eighties to primarily quantitative empirical evaluation in the present. Figure 2 shows how the eighties included papers both with and without evaluation, where in 2000 and 2006, virtually all papers

included evaluation and informal evaluation has almost disappeared in 2006[1].

As illustrated in figure 2, analytical evaluation has never been widespread in CHI and the change in proportion is more likely due to the small sample than an actual trend; no more than three papers using analytical evaluation were found in any year. Most analytical evaluations were found to be analyzing log data or GOMS analyses of graphical interfaces. We return to this finding in the discussion.

Papers whose contribution was primarily an evaluation method were classified in their own category. We recognize some evaluation methods can be evaluated (and were in many cases, in others cases not), but our rationale for excluding them is that they were meta-papers in our survey and as such would not logically fit into our analysis. These make up between zero (2006) and fourteen (1983) percent of the papers with a clear linear decrease through the years. These papers' contribution is often the creation and promotion of a usability evaluation technique. CHI authors in 1983 presented methods such as evaluation technique using a mockup user interface [21]. In the early 90s, se saw the 'Damaged Merchandise' debate when we as a community (although published outside the CHI conference) debated the validity of discount usability methods [14]. Perhaps, in partial response to this debate, even fewer papers addressed usability evaluation techniques. In 2000 only two papers looked at evaluation methods and in 2006, no papers were found in this category. The most prominent of the evaluation types in our sample, however, was found to be empirical evaluation, qualitative as well as quantitative. We now turn to discuss these types in detail.

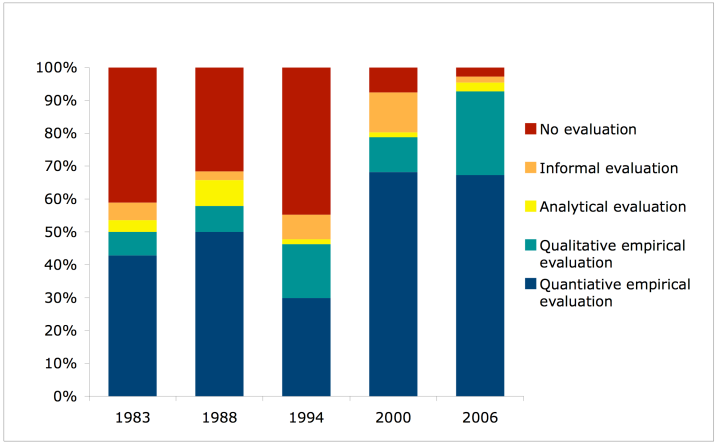

Figure 2: Indexing of different types of evaluation styles through five selected years. Note that papers are indexed on the basis of the *main* type of evaluation, e.g. many evaluations combine quantitative with qualitative measures.

[1] Note that the figure only includes papers presenting technologies or technical concepts; hence, overview papers and papers specifically on evaluation methods are not included.

Changes in Empirical Evaluation

Quantitative Empirical Evaluation

A classic evaluation type in HCI is the quantitatively conducted experiment. Figure 3 shows how the presence of that evaluation type has not changed much though the years. The distribution goes up and down, with 1994 and 2006 as slightly diverging years, but CHI is still mainly doing quantitative evaluation, with about 12 percent of these supplementing the evaluation with qualitative measures. In 2006 over half of the papers (61 out of 118) presented this type of evaluation. These experiments are often conducted as task-oriented within group experiments comparing the authors' new technology or system to an older one. Thorough analysis is then conducted, often using statistical significance tests such as ANOVA to provide evidence for the claims.

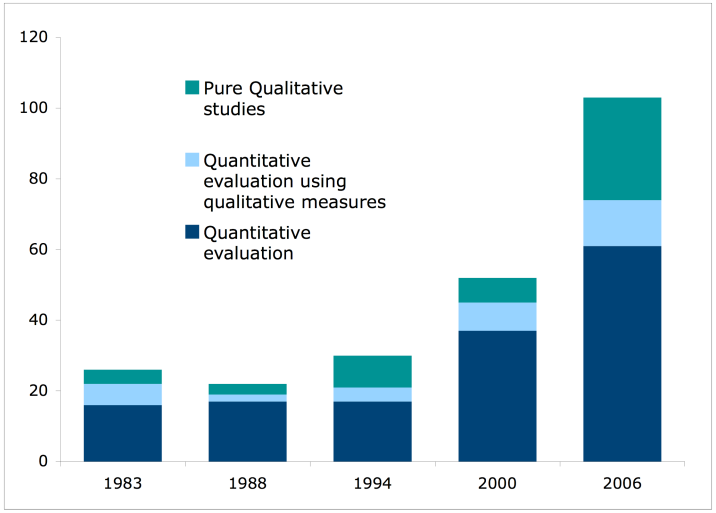

Figure 3: Empirical studies according to research method.

Qualitative Empirical Evaluation

The emergence of CSCW in the mid-eighties and its 'mingling' with CHI is witnessed by in an increase of qualitative studies of present technologies or technology settings, often conducted using ethnographic methods, evident in our survey from the nineties. Although the numerical change in the proportion of papers, which are qualitative, does not seem to be very dramatic, as illustrated in Table 1, when taking a closer look at the content of the specific studies, the change in focus becomes clear. For example, Mantei and Haskell looked at a user's first experiences with home microcomputer applications in 1983 [26], but by 1994 papers were concentrating on a more diverse set of software technologies, such as Kidd's study [22] of how technology supports the everyday job of knowledge workers. Pure qualitative studies then decrease in 2000 but increase again in 2006, where researchers venture into new fields, for example emergency services [23] and everyday gameplay [2].

1983	1988	1994	2000	2006
7%	5%	11%	7%	14%

Table 1: Proportion of papers providing a qualitative, often explorative, study of existing technologies or technical environments.

Blending Qualitative and Quantitative Measures

Another common method is the blending of qualitative and quantitative approaches. Practitioners of this approach primarily use a quantitative approach, supplemented with a few qualitative measures. For example Salvucci and Anderson evaluated their gaze-

based interface by measuring how many tasks the participants perform correct and the time they take (quantitative data) before interviewing them informally about their experience and strategy for use (qualitative data) [34]. The interesting trend here is that although it would be logical to see this increase through the years alongside pure qualitative evaluation, the proportion stagnates. In fact 1983 sees the biggest proportion of blended studies (27 percent), 2000 and 2006 both see only 17 percent blended studies. We want to draw attention to this as a potential weakness in evaluation methodologies and return to this finding in the discussion.

Changes in Subject Selection

One of the goals of this survey was also to take a critical look at the potential users we evaluate our newly developed technologies and technical concepts with, both in terms of numbers and type of subjects. If one cynically assumes that a primary role of evaluation for systems builders in CHI is to provide statistical significant evidence of the desirability or efficiency of the author's system over previous work, one would expect that studies would rely on statistical power, and thus large sample size. By taking a closer look at the number of subjects used in quantitative and qualitative evaluation respectively, we find an interesting trend as illustrated in figure 4. The median number of subjects in the quantitative empirical studies has decreased over time, and the median number of subjects in qualitative studies seems to have increased (although less clear a trend).

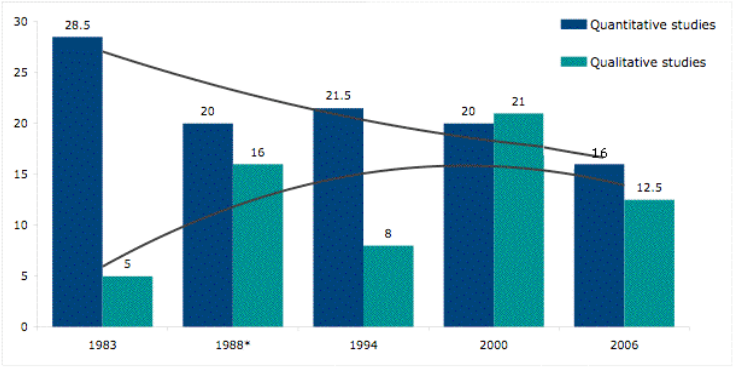

Figure 4: Median number of participants in the empirical evaluation studies. *Note that the median for 1988, qualitative evaluation only includes two studies, one using 29 subjects and one using three; none of the other qualitative studies that year mention the number of subjects.

Despite fluctuations, in median[2], empirical quantitative evaluations are clearly using a smaller number of participants than they were 24 years ago. Early studies often used between fifty and hundred subjects in their experiment, where it is more common to see experiments using less than twenty participants in 2006. On the other hand, the number of participants in purely qualitative studies has generally increased.

One possible interpretation of the changing numbers could be that the two types of research are slowly coming together and borrowing approaches from each other. However, our finding that the proportion of

[2] We are using the median of number of subjects instead of average because it is traditionally more representative of a large variety of data and counter a single outlier.

quantitative evaluations using qualitative data is still quite low contradicts this. Alternatively, one could contribute the decrease in participants to the increasing pressure to provide a supportive quantitative evaluation. Qualitative evaluation, on the other hand, has transition from in-depth studies of a few key informants to studying larger populations. A potential cause is pressure from more statistically oriented colleagues, who will potentially review the paper. Yet such an approach threatens the integrity of an ethnographic or case study approach.

The percentage of papers containing evaluation has increased from just over half of the papers in 1983 to 97 percent of the papers in 2006. It is interesting to speculate how the overall increase in the amount of evaluation, combined with an overall decrease in subjects may have lowered the quality of the evaluation. This brings us on to the next level of analysis, looking at the type of subjects used in quantitative evaluation.

STUDENT-COMPUTER INTERACTION

Many traditional psychology experiments often use students as their main population. Psychologists justify this by the fact that students, despite their youth and education, are representative of the population at large for many studies — visual perception, rudimentary understanding of memory, etc. Not surprisingly, HCI research followed suit. In early HCI, students, while junior, could reasonably be studied in lieu of more senior computer scientists for whom machines were designed. Further, students in their first year could participate as novice users without prior computing experience. Now that computers are more widespread and many applications are targeting a diverse set of people, students have too much computer savvy to be representative of the entire spectrum of novice to expert users. Moreover, them being in an educational setting and used to learning new things makes them unusual in terms of ability to learn. Despite these significant discrepancies with 'typical' users, half of the studies in our sample conducted their experimental work using either undergraduates or graduates students[3].

Two significant variations should be noted. First, studies of technologies intended for specific target populations, such as the elderly or disabled (e.g. [33]) do not follow this trend. Second, in the eighties and to a certain extend early nineties, it was not common to specify the type of participants used in evaluation. Once it became common to specify general information about the participants, we still see papers relying primarily on students — 57 percent of the papers mentioning this information in 2000, dropping to 48 percent in 2006. On the one hand, the increased discussion of participant type is indicative of an increasing openness in the evaluation process, on the other, the percentage of student participants is unjustifiably high.

Although students are a great resource for many researchers, they cannot be considered as representative of any general population, their youth and active learning environment mean they are quick

[3] This does not include studies that do not specify their population. Ethnographic studies and case studies are not included either, since they per definition use subjects of relevance to the environment in question.

learners and often have more experience with computers and technology (in particular the computer science students often used) than the average population. So lest we wish to change our field's name to student-computer interaction we should make effort to find more representative participants.

LACK OF FEMALE PARTICIPANTS

In addition to the high number of experiments using students, we found that many studies failed to use a gender-balanced sample. Many fail to mention the gender distribution altogether, particularly the early years.

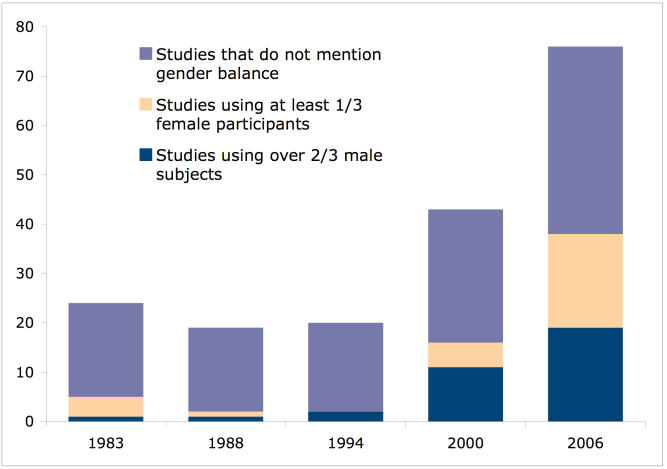

Figure 5: proportion of female participants in empirical quantitative evaluations though the selected years.

There are few exceptions of studies having more female than male subjects (e.g. [9, 12]), but in general 25 percent of the empirical quantitative studies in 2006 use an unacceptable low proportion of females. Figure 5 shows the variation through the years. As illustrated, similar to our finding about students, most studies before 2000 did not provide gender information either. Some studies, however, mention other types of information; one set of authors for example found it more important to mention their student participants' SAT score than the gender balance [8], although their research did not depend on subjects having specific cognitive skills. Again, in 2000 we see a rise in information about participants; however, studies from 2000 and 2006 clearly do not include many women. We will return to this point in the discussion.

Discussion

We have in this paper pointed to both trends in changing use of evaluation methods in CHI papers. We want to address the three questions that we started out by raising, and try to put the results into the larger context of our field. We hope that the findings from our survey will help our community understand how it has evolved, and understand how the history of evaluation will continue to influence the future.

Evaluation over the last 24 years

The CHI community has grown to be evaluation centric. It is no longer accepted practice to submit papers without evaluation. At the same time, use of evaluation has become a political tool for a paper's acceptance. Our study shows how our community is increasingly relying on empirical studies to validate our work (from less than half of papers including empirical evaluation in the early eighties to the vast majority today). At the

same time we found that subject numbers are at an all time low and that the general diversity of subjects is weak. Consequently, we might ask ourselves if we are advocating discount evaluation without its label. Although the papers, on the whole, have evolved from five page descriptions of new interface technologies to ten page reports of studies of complex systems, there is no need to underestimate the contribution of the former type of paper. While the premise that some research should not require evaluation is no doubt contentious, by valuing empirical research as highly as we do, we might lose ground breaking papers that cannot provide clear empirical evidence that their technology is an improvement over an older one.

We need to recognize that in a search for credibility within our community, creators of new technologies are likely to rely on mainstream evaluation methods. For those whose primary focus is on creating new technologies, for whom evaluation is a secondary concern, the continual change in accepted evaluation practice presents an obstacle for their work in a requirement to master new approaches. The existence of a norm in our community reinforces the legitimacy of established approaches, and presents a challenge for those creating new evaluation methods. Our field then also runs the risk of becoming 'stiff' in its methods; instead of embracing a number of different evaluation methods, some possibly more appropriate to the research in question, we focus on the method of the status quo. This brings us on to the next finding, the finding that CHI does not embrace analytical evaluation.

When considering the emphasis on evaluation in CHI it is surprising how little research includes analytical evaluation. These approaches require respect of expert opinion, as opposed to the neutral statements of a participant. A potential weakness is that expert opinion raises the potential of bias, especially if validating ones own work. From the observation that CHI has rarely included this type of evaluation, we reason that it is viewed as a less valid approach. We can therefore not help but wonder, with the Damaged Merchandise debate less than ten years behind us, if we as a community are simply avoiding having to answer the questions this debate raised about validity, by not using these approaches. Finally, one reason might be that it is more 'interesting' to evaluate with potential users than to analyze tons of logs or perform a Heuristic Evaluation analysis on yet another new user interface. The consequence is that analytical evaluation is likely to always be a minority in CHI, however, one that adds to diversity of evaluation methods.

The diversity of evaluation methods also comes into play with the decrease in papers presenting evaluation methods themselves. Partly, this is indicative of a spilt between practitioners and academics. We have seen a diminishing role of practitioners in CHI, an issue that has been discussed the last several years [1, 31]. Within industry analytic evaluation techniques are relied heavily upon, and yet our study has showed they are no longer commonly discussed or used within the CHI community. We have to wonder why approaches like Heuristic Evaluation and Cognitive Walkthrough, which are so commonplace in industry are no-longer acceptable practice within CHI. Perhaps, practitioner techniques are not included because they evolve to accommodate demands of business which are viewed somehow un-scientific. Whatever the reason, we have to ask why techniques developed to accommodate

business needs are not relevant to CHI, given that businesses provide an important gateway controlling how technology is distributed to consumers. We alluded to many potential reasons, above but without addressing these issues the practitioner/academic division is only likely to widen. While the advent of the online commerce for example, required a rapid response that we saw in the push towards discount evaluation, new computing trends—ubiquitous, social or tangible computing—do not present such urgency; yet there is a considerable body of recent work debating a significant shift in HCI evaluation, which includes a movement towards evaluation experience [25, 27, 35], work discussing the role of ethnography in HCI [10], and discussion of how the need for evaluation impacts the ability to present design work at CHI [39]. Much of this work has been highly contentious, and indeed much of it presented at HCI venues outside of CHI. Regardless, the proportion of CHI work that discusses the role of evaluation is very small, which makes us question whether evaluation is in fact responding dynamically to the radical changes in technological innovation. This causes us to reflect on the role of evaluation in our community, if it is converging into a set of rules to be followed and not questioned. Now, having reflected on the context and potential limitations of having a clear form of accepted practice, we move to discussing empirical evaluation itself.

Empirical Evaluation and its Importance in CHI

As was clear from our analysis of evaluation methods, quantitative evaluation methods was a 'winner'. This type of evaluation has a long tradition in CHI, older than the conference itself and is a great tool for validation. A positive trend that we observed was a recent increase in qualitative evaluation studies, studies often taking place over longer time and using multiple sets of inquiry methods. This is a trend that illustrates how evaluation is not just a validation tool. It provides us with indications of user appropriation and contextual fitting of the technology in question. Qualitative and situated evaluation also informs the next iteration of technology design and how technology will be used in practice. As we branch away from office applications to applications for the home, the subway, the grocery store, and the places in between, techniques that adequately interpret context will be increasingly important. Ethnographies, for example, can provide insight into situated technology use and a social setting but it can say little about a systems objective efficiency. To our toolbox of techniques, we need ways of measuring the quality of user experience, in addition to the usability of the technology itself.

In this spirit it was surprising to see so few evaluations using blended approaches. Quantitative evaluation can only address a certain set of issues and with increasingly complex settings of technologies, we need an increasingly encompassing set of methods. Gaver for example, warns that the transfer of the computer from the office to the home will bring with it workplace values such as efficiency and productivity at the expense of possibilities for exploration and enjoyment [Gaver et al. 2004]. Technology should support a multitude of goals, office efficiency only being one of many.

Subject Selection and Limited Diversity

Finally we return to our last part of the data, looking at the trends with regards to subject selection in CHI.

The changes in number of subjects used in empirical evaluation is interesting, however, it is perhaps more indicative of changes in evaluation rather than improvement or decrease in the quality of evaluation. Quantitative research will by nature always rely on numbers and by using smaller numbers of subjects, researchers are mainly making their own job harder. It becomes more difficult to convince the audience that their technology is in fact better than another or that the claimed factors can be generalized. Qualitative studies do not have to be generalizable to the same extent, making it possible to use fewer participants. Where the few qualitative studies presented in the eighties were in-depth evaluations over long time with few participants, the prevalence of these studies bring with them adaptations of these methods, using more participants, but being less detailed. The advantage is the emergence of more studies of this type, but it is important to note that an increase in numbers does not make up for a lack in depth. Besides change in subject numbers, we found an overrepresentation of both students and male participants in the quantitative empirical evaluation studies. Although both are important, focusing on students instead of representative users has fairly obvious consequences; we therefore want to focus our discussion on under-representing the number of women tested, which has more insidious results.

Relating to the issue of more encompassing evaluation are feminist studies of technology. Through the last couple of decades we have seen an increase in not just feminist critiques of technology but also empirical studies of gender issues within different technologies [16]. One key argument is that the inherent male bias of technology is in part caused by women's lack of involvement in the design of technologies such that these are shaped by male power and interests [38 in 16]. Feminist studies argue that technologies are created in the context of male culture and embody certain assumptions of female life. This in return means that women are alienated by technology and define their femininity in terms of rejection of technology rather than encompassing it [37 in 16] An obvious way of countering this trend would be to include more (or at least a proportion similar to the user population) females in technology evaluation. To date this has not occurred in HCI, and we are perhaps perpetrating the design of a next generation of gender biased technology. The evaluation of computing technologies is done primarily with male participants, leading to support the cyclic nature of technology development and use as male orientated.

Conclusion

In reflecting on the last 24 years of HCI evaluation we need to recognize how the suite of techniques has evolved and will continue to do so in response to new technologies. In our analysis of evaluation approaches we found numerous nuances and trends that are important to be aware of if we as a science and practitioner community want to evolve. The answer might not be as straightforward as it seems, such as increasing diversity of experiment subjects; instead it is important that we acknowledge the need for reflection on these topics. One view is to see the developments within CHI as the natural development of a new field. Newman acknowledged already in 1994 that HCI is not a science that provides the same outcome and contributions as a traditional engineering discipline [30]. A young field naturally becomes more 'scientific' through time, using more rigid methods and less

alternative techniques, because of an increasing consensus of accepted methods. The overall consequence, as indicated by our survey, is a set of accepted evaluation methods that have to be followed in order for the research to fit into the field. By pointing to this consequence we propose to review these methods and open up for new developments in evaluation. After all, the types of evaluation tools appropriate for studying input devices like mice, have evolved to encompass other types of input devices; these tools will continue to evolve in order to encompass the need to evaluate next generation socially situated ubiquitous technologies for both men and women. We conclude in the spirit of Burns who did not think much of our approach to evaluation:

The best-laid schemes o' mice an' men
Gang aft agley[4],
An'lea'e us nought but grief an' pain,
For promis'd joy! [Burnes, 1785]

Acknowledgements

We thank Jofish Kaye for assistance with indexing of papers, Saul Greenberg for comments and encouragement, as well as Mike Kuniavsky for practitioner insights. Lastly we are grateful for the careful critique of reviewers.

References

1. Arnowitz, J. and Dykstra-Erickson, E. 2005. CHI and the practitioner dilemma. Interactions 12, 4 (Jul. 2005), 5-9.
2. Bell, M., Chalmers, M., Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown, B., Rowland, D., and Benford, S. 2006. Interweaving mobile games with everyday life. *In Proceedings of CHI '06*. ACM Press, New York, NY, 417-426.
3. Benford, S., Crabtree, A., Reeves, S., Sheridan, J., Dix, A., Flintham, M., and Drozd, A. 2006. Designing for the opportunities and risks of staging digital experiences in public settings. In *Proceedings of CHI '06*. ACM Press, New York, NY, 427-436.
4. Burnes, Robert. "To a Mouse" In *Kilmarnock Volume*, 1785.
5. Calingaert, P. 1967. System performance evaluation: survey and appraisal. Communication of the ACM 10, 1 (Jan. 1967), 12-18.
6. Cheriton, D. R. 1976. Man-machine interface design for timesharing systems. In Proceedings of the Annual Conference (Houston, Texas, United States, October 20 - 22, 1976). ACM 76. ACM Press, New York, NY, 362-366.
7. CHI website, guide to papers: http://www.chi2007.org/submit/papers.php
8. Corbett, A. and Trask, H. 2000. Instructional interventions in computer-based tutoring: differential impact on learning time and accuracy. In *Proceedings CHI '00*. ACM Press, New York, NY, 97-104.
9. Dey, A. K. and de Guzman, E. 2006. From awareness to connectedness: the design and deployment of presence displays. In *Proceedings of CHI '06*. ACM Press, New York, NY, 899-908.
10. Dourish, P. 2006. Implications for design. In *Proceedings of CHI '06*. ACM Press, New York, NY, 541-550.
11. Ducheneaut, N., Yee, N., Nickell, E., and Moore, R. J. 2006. "Alone together?": exploring the social dynamics of massively multiplayer online games. In

[4] Often translated into modern English as "often go awry".

Proceedings of CHI '06. ACM Press, New York, NY, 407-416.

12. Egido, C. and Patterson, J. 1988. Pictures and category labels as navigational aids for catalog browsing. In Proceedings of CHI '88. ACM Press, New York, NY, 127-132.
13. Gaver, W. W., Bowers, J., Boucher, A., Gellerson, H., Pennington, S., Schmidt, A., Steed, A., Villars, N., and Walker, B. 2004. The drift table: designing for ludic engagement. In *CHI '04 Extended Abstracts*. ACM Press, New York, NY, 885-900..
14. Gray, W. D. & Salzman, M. 1998. Damaged Merchandise? A Review of Experiments that Compare Usability Evaluation Methods, HCI, 13(3), 203-261.
15. Greenberg, S. and Witten, I. H. 1988. How users repeat their actions on computers: principles for design of history mechanisms. In *Proceedings of CHI '88*. ACM Press, New York, NY, 171-178.
16. Grint, K. and Gill, R. The Gender-Technology Relation. Contemporary Theory and Research. Taylor and Francis, London, UK, 1995.
17. Grudin, J. 2005. Why CHI fragmented. In CHI '05 Extended Abstracts on Human Factors in Computing Systems (Portland, OR, USA, April 02 - 07, 2005). CHI '05. ACM Press, New York, NY, 1083-1084.
18. Grudin, J. 2006. Is HCI homeless?: in search of inter-disciplinary status. Interactions 13, 1 (Jan. 2006), 54-59.
19. Haggett, A. G., McFadden, J. R., and Newsted, P. R. 1981. Naive user behavior in a restricted interactive command environment (abstract only). In Proceedings of the Joint Conference on Easier and More Productive Use of Computer Systems. Human interface and the User interface - Volume 1981. ACM Press, New York, NY, 139.
20. Hammond, N., Jørgensen, A., MacLean, A., Barnard, P., and Long, J. 1983. Design practice and interface usability: Evidence from interviews with designers. In *Proceedings of CHI '83*. ACM Press, New York, NY, 40-44.
21. Jacob, R. J. 1983. Executable specifications for a human-computer interface. In *Proceedings of CHI '83*. ACM Press, New York, NY, 28-34.
22. Kidd, A. 1994. The marks are on the knowledge worker. In *Proceedings of CHI '94*. ACM Press, New York, NY, 186-191
23. Kristensen, M., Kyng, M., and Palen, L. 2006. Participatory design in emergency medical service: designing for future practice. In *Proceedings of CHI '06*. ACM Press, New York, NY, 161-170.
24. Mackay, W. E. 1998. The CHI conference review process: writing and interpreting paper reviews. In CHI '98. ACM Press, New York, NY, 376.
25. Mandryk, R. L., Atkins, M. S., and Inkpen, K. M. 2006. A continuous and objective evaluation of emotional experience with interactive play environments. In *Proceedings of CHI '06*. ACM Press, New York, NY, 1027-1036.
26. Mantei, M. and Haskell, N. 1983. Autobiography of a first-time discretionary microcomputer user. In Proceedings of CHI '83. ACM Press, New York, NY, 286-290.
27. Monk, A., Hassenzahl, M., Blythe, M., and Reed, D Funology: designing enjoyment, CHI '02 extended abstracts ACM Press, New York, NY.
28. Morse, A. 1979. Some principles for the effective display of data. In Proceedings of the 6th Annual Conference on Computer Graphics and interactive Techniques (Chicago, Illinois, United States, August 08 - 10, 1979). SIGGRAPH '79. ACM Press, New York, NY, 94-101.
29. Nardi, B. A. and Johnson, J. A. 1994. User preferences for task-specific vs. generic application software. In Proceedings of the CHI 1994, 392-398.

30. Newman, W. 1994. A preliminary analysis of the products of HCI research, using pro forma abstracts. In Proceedings of CHI '94. ACM Press, New York, NY, 278-284.

31. Nielsen, J. 1989. Usability engineering at a discount. In Proceedings of the Third international Conference on Human-Computer interaction on Designing and Using Human-Computer interfaces and Knowledge Based Systems, Elsevier Science, New York, NY, 394-401.

32. Parush, A. 2006. Toward a common ground: practice and research in HCI. Interactions 13, 6 (Nov. 2006), 61-62.

33. Petrie, H., Hamilton, F., King, N., and Pavan, P. 2006. Remote usability evaluations With disabled people. In Proceedings of CHI '06. ACM Press, New York, NY, 1133-1141

34. Salvucci, D. D. and Anderson, J. R. 2000. Intelligent gaze-added interfaces. In Proceedings of CHI '00. ACM Press, New York, NY, 273-280.

35. Sengers, P. and Gaver, B. 2006. Staying open to interpretation: engaging multiple meanings in design and evaluation. In Proceedings of DIS '06. ACM Press, New York, NY, 99-108.

36. Soloway, E., Ehrlich, K., and Bonar, J. 1982. Tapping into tacit programming knowledge. In Proceedings of the 1982 Conference on Human Factors in Computing Systems (Gaithersburg, Maryland, United States, March 15 - 17, 1982). ACM Press, New York, NY, 52-57.

37. Turkel, S. 1988. "Computational Reticnece: Why Women fear the Intimate Machine", In Kramarae, C. (ed.) Technology and Women's voices. NY, Routeledge and Kegan Paul.

38. Wajcman, J. (1992) Feminist Theories of technology, Paper presented at workshop on The Gender-Technology Relation, CRICT, Brunell University 16-17 Sep.

39. Wolf, T. V., Rode, J. A., Sussman, J., and Kellogg, W. A. 2006. Dispelling "design" as the black art of CHI. In *Proceedings of CHI '06*. ACM Press, New York, NY, 521-530.

40. Wulff, W. and Mahling, D. 1990. An Assessment of HCI: Issues and Implications. SIGCHI Bulletin July 1990. 22 (1) ACM Press, New York, NY, 80-87.