Foundations and Trends® in
Human–Computer Interaction
Vol. 5, No. 4 (2011) 299–373

DOI: 10.1561/1100000043

now
the essence of knowledge

Some Whys and Hows of Experiments in Human–Computer Interaction

Kasper Hornbæk

Department of Computer Science, University of Copenhagen, DK-2300, Copenhagen S, Denmark, kash@diku.dk

Abstract

Experiments help to understand human–computer interaction and to characterize the value of user interfaces. Yet, few intermediate guidelines exist on how to design, run, and report experiments. The present monograph presents such guidelines. We briefly argue why experiments are invaluable for advancing human–computer interaction beyond technical innovation. We then identify heuristics of doing good experiments, including how to build on existing work in devising hypotheses and selecting measures; how to craft challenging comparisons, rather than biased win–lose setups; how to design experiments so as to rule out alternative explanations; how to provide evidence for conclusions; and how to narrate findings. These heuristics are exemplified by excellent experiments in human–computer interaction.

Contents

1

Introduction

This work began as an attempt to answer a colleague's question. For some time I had insisted that we run experiments on a new interaction paradigm that we had been working on. My colleague had asked for papers that would convince him why we should do experiments at all. He also quickly asked for papers that explained how to do those experiments, seeing my expression of disbelief after the first question. I was unable, however, to give him entirely satisfactory references: this forms the background for the present work.

A fair number of papers describe how to do experiments in human–computer interaction (HCI). For instance, Landauer [86] gave a classic discussion of research methods in HCI, including valuable advice on statistical analysis and reporting. Blandford and colleagues [14] discussed how to plan, run, and report experiments in HCI, and presented an illustrative case study. Recently, Lazar et al. [90] published a book on research methods in HCI that included several chapters on designing and reporting experiments. Also, a number of papers review experimentation on topics closely related to HCI, including information retrieval [81], information visualization [21], and text editing [118]. More generally, a host of literature relevant to the design of

experiments exists in the field of psychology [95, 122], sociology [142], and ergonomics [34].

Why, then, another paper on experiments in HCI? First, the above papers focus little on the questions that arise even when you understand the basics of experimental logic, the distinction between independent and dependent variables, and the concerns in ensuring statistical conclusion validity. Second, many of the papers referenced above focus little on the specific difficulties arising from experimenting with interfaces and interactions. Third, while papers on specific topics are helpful, they de-emphasize that many areas of HCI face similar questions about why and how to do experiments.

We consider an experiment "a study in which an intervention is deliberately introduced to observe its effects" [127], p. 12. The intervention may be of a variety of kinds; in HCI it is often a technology, but could be kinds of training, user group, use situation, or task. We follow common practice by designating the intervention as a level of an independent variable, or as a treatment, or as a condition. The effects of the intervention are measured as dependent variables. In HCI they will often include measures of the usability of the technology. Hypotheses are statements that connect variation in independent variables to expectations about variation in the dependent variables. Another defining characteristic of experiments is that they attempt to deal with other factors besides the independent variable that influence the situation under study, and thus potentially affect the dependent variables [42]. This may happen, for instance, by controlling such factors, holding them constant, or distributing them randomly across levels of the independent variable. Finally, it is typical of experiments that the situation under study is created or initiated by the experimenter [42]. Figure 1.1 shows an outline of these components. Note that the above definition excludes the understanding implied in some common usages of the word experiment, including that of "trying something new" or "an innovative act or procedure".

The logic underlying experiments is tied to pioneering work in the renaissance, in particular by Galileo Galilei and Francis Bacon. Later, John Stuart Mill refined thinking about experiments by his Joint Method of Agreement and Difference. The key idea is that effects occur

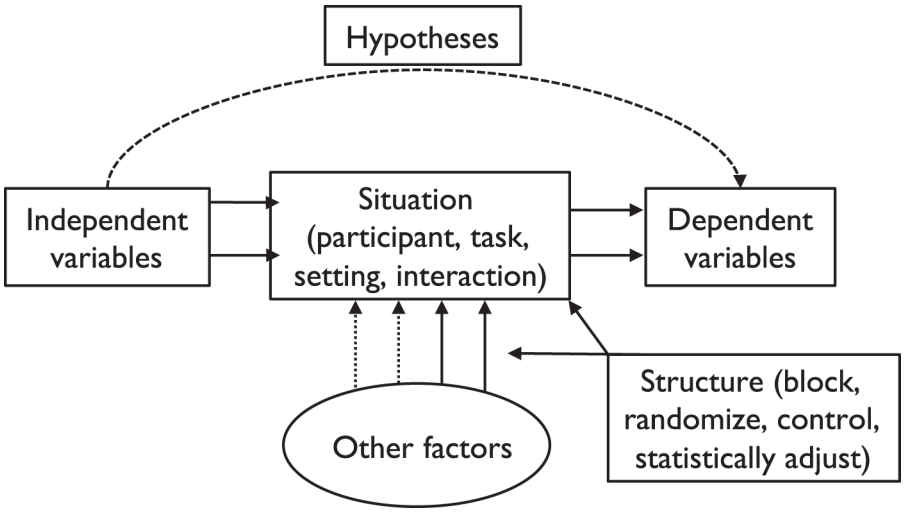

Fig. 1.1 Typical components of experiments in human–computer interaction.

with their presumed causes and that any difference between outcomes may be used to attribute causes; this idea is directly reflected in the above definition of experiment. Bunge [15] and Shadish et al. [127] further discuss the logic underlying experimentation and its historical development.

2

Why Conduct Experiments?

Experiments in HCI may be undertaken with a variety of aims. For instance, experiments may be used to evaluate existing or new interfaces, establish guidelines and standards, discover scientific principles, understand people's use of technology, and test models of performance. The present work concentrates on experiments where the levels of the independent variables of key interest are user interfaces (UIs).

2.1 Reasons for Experiments

At a general level, it may be argued that experiments are necessary in HCI to validate the technologies that we develop. The point has repeatedly been made that computer scientists validate their technologies too infrequently [47, 135, 146]. Zhai et al. [147] argued that while empirical evaluation has its problems, it is far superior to merely asserting that a technology is valuable. Newman [106] compared abstracts of engineering journals with abstracts of HCI proceedings. In engineering, papers more often presented validations of modeling techniques or technical solutions; in HCI, papers often presented radical solutions, that is, solutions building on or extending previous work. Validations in HCI were much more rare than in engineering. Independently of

whether the engineering profession is a viable ideal for HCI, Newman's data suggest that we were — and perhaps still are — much better at proposing new technologies than at validating them.

A second reason for doing experiments is that they are particularly valuable for answering some research questions in comparison to alternative methods. Specifically, experiments allow us to investigate use of technology without deploying it, to compress time and thereby study infrequent phenomena, to study interaction that would otherwise be unethical or impossible to obtain data about, to control external factors, and to collect fine-grained data. It has been argued that all research methods, including experiments, are valuable for particular purposes. In a paper on research methods that has been influential in HCI, McGrath [96] stated that "all methods have inherent flaws, though each has certain potential advantages" (p. 154). Weber [139] made a compelling argument for method pluralism in asserting that "excellent researchers simply choose a research method that fits their purposes and get on with the business of doing their research" (p. xi) and Firebaugh [37] wrote "Let method be the servant". Platt [117, p. 351] presented a similar argument:

> Beware of the man of one method or one instrument, either experimental or theoretical. He tends to become method-oriented rather than problem-oriented. The method-oriented man is shackled; the problem-oriented man is at least reaching freely toward what is most important.

Thus, independently of the vogues and waves of HCI, the argument made here is that experiments form an essential part of our methodology toolbox.

A third reason for experiments in HCI is to battle what Landauer [86] called the "egocentric intuition fallacy". Landauer argued that we overestimate how well we can discover the mechanisms behind our behavior and feeling of satisfaction. That makes it difficult to rely on judgments about what works and does not work in an interface. At the same time, we underestimate the extent to which we differ from other people with respect to behavior and preferences. This is particularly

problematic in HCI, where a number of the technologies that we propose will have been developed and iteratively refined by ourselves or by close collaborators. Experiments help overcome this fallacy.

A common criticism of experiments is that they are hard to generalize because they are artificial and narrow in scope. To this, several answers are possible. McGrath [96] would argue that precision and control in experiments come at the expense of realism. In laboratory experiments, the situation under study would not occur without the experimenter; participants' motivation is thus a key difference between experiments and, say, field studies. Another answer is that experiments may not have generalization as their goal. Rather, they may be used to show that something may happen (what Greenberg and Buxton [50] called "existence proof"), to test generalizations (rather than making them), or to understand the processes involved in a particular phenomenon [100]. Still another answer is that the criticism that experiments do not generalize to field settings may be invalid (or at best, requiring an empirical answer). Campbell [18], for instance, argued that in organizational psychology there is probably little difference between the results of lab and field studies; Anderson et al. [6] argued that the effect sizes found in laboratory and field research correlate strongly. In contrast, Gerber and Green [45] argued that there are many differences between the results of lab and field studies. We will not attempt to settle this complex discussion. Rather, because of the criticism of experiments, we simply wanted to argue that in some circumstances experiments are a useful and valid research approach.

2.2 Alternatives to Experiments

The previous section makes it clear that there are good arguments for using experiments to answer some research questions. Following the arguments attributed to McGrath [96], experiments — like any other research method — have many limitations. Therefore, we briefly discuss some alternatives and their relation to experiments. An in-depth discussion of those alternative research methods is outside the scope of this work, see instead any broadly covering book on research methods (e.g., [90, 119]).

In HCI, a couple of situations stand out where one should not do an experiment, although many more may be identified for specific research questions. One such situation is when the researchers are interested in the uptake or appropriation of technology. Here, users' choices and motivation in using (or not using) technology are key factors, as is the realism of their use. Often the aim is also to describe in rich ways how users engage with technology. In such cases, an experiment is often not the right method. Instead, one should think first about using a method that emphasizes realism and/or rich description, such as a field study or a qualitative interview.

A related situation is when the phenomenon of interest is not sufficiently well understood to set up an experiment. This echoes the sentiment that "premature experimentation is a common research sin" [127, p. 99]. Such a situation would occur if candidates for independent variables cannot be identified or if more than one level of that variable (say, an interface to compare an innovation against) cannot be imagined. In my view, the latter sometimes occur for user interface developments whose main contributions are technical or engineering. Experiments are often attached to reports of such developments, even if their outcome is mostly given in advance or their quality so low as to distract readers from the main contribution. I believe that it is important to appreciate — both for authors and reviewers — that many technical/engineering contributions are wholly adequate for publishing without an evaluation.

A final situation where experiments are unneeded occurs when the aim is to drive development of an interface. In such cases, the rigor of experiments is often not needed. Instead one may do usability studies with few participants, possible studying only one interface condition. Such a study may give plenty of information on how to improve an interface (see [108] for advice on how to run such a study). Alternatively, one may engage in longer-lasting iterative development with prospective users, such as in long-term case studies [129]. Also in those situations, the evaluations mainly serve to drive development and understand utility of an interface to users; their goal is not to make rigorous inferences about differences among conditions.

3

How to Conduct Good Experiments?

Good experiments are hard to do. Before detailing the tactics of experiments, we briefly discuss how to think about experiments at a slightly higher level of abstraction. We first emphasize the need to be clear about the research question or, put in another way, the purpose of running the experiment. Then, we discuss some high-level heuristics that in our view help doing good experiments.

3.1 Finding a Significant and Interesting Research Question

Any consideration of how to do an experiment must depart from the research question one wishes to address. Finding out what to study is outside the scope of this work; see for instance Campbell et al. [19] or McGuire [97] for inspiring ways to think about and develop research questions. Here we assume that the research question being asked is significant and interesting. To ensure this, one can think over two potential objections to a finished and written up experiment: (1) "so what" and (2) "no surprises" [37]. The "so what" objection suggests that imagined results of an experiment should be interesting and nontrivial; they should matter to theory or practice. Even if running and analyzing the experiment proceeds as imagined, will people find it interesting? Will it

add to our understanding of HCI in important ways? Sometimes, this objection is voiced by reviewers as "this is not significant", meaning that while the findings are novel and valid, they do not add to the research literature in a substantial and important manner.

The "no surprises" objection suggests that results should add to or depart from what we already know; they should not be predictable given earlier studies. As mentioned earlier, one should not do an experiment if the results are clear in advance. For instance, if a simple predictive model shows a user interface superior to another or if a technology is obviously superior to an alternative then an experiment does not have the possibility to surprise us. Sometimes, of course, new technologies, use situations, or user groups may make it hard to know if earlier findings or theories apply. The "no surprises" objection may be raised both because the experimental setup is biased (we will discuss how to avoid this later) and because the results are easily predictable from the literature (we will also discuss how to avoid this later). Most importantly, both of these objections can and should be considered before deciding to run an experiment.

Finding significant and interesting research questions requires a solid grasp of the literature. However, it is out of the scope of this work to discuss how to identify and retrieve relevant earlier research (see for instance Cooper [28]).

3.2 Some Heuristics for Good Experiments

The type of advice that the present work aims to give must necessarily be heuristic and personal. Heuristic means that we give only weak guidance at a level of abstraction to be fleshed out, traded-off, and creatively applied; personal means that this guidance to some extent is a matter of taste and style. Neither of these characteristics means that there are no wrong decisions in experimental design, nor that anything goes. Rather they suggest that some decisions in experimental design are really complex, requiring creative ideas and difficulty choices. Table 3.1 summarizes heuristics on how to conduct experiments. The next sections will discuss in detail how to use these heuristics to design, run, and report experiments.

Table 3.1. Heuristics for conducting experiments.

Heuristic	Explanation	How to?
Be focused	Focus the experiment through a clear research question that drives the design and interpretation of results.	Let the research question prescribe methods and measures; simplify the design; formulate hypotheses when feasible; highlight contribution; produce few "ticks".
Use previous work	Build on previous work in designing, running, and reporting experiments.	Motivate hypotheses by data and theories; use validated ways of measuring; replicate earlier findings; show importance over prior work.
Do strong comparisons	Make a challenging and multifaceted comparison, and prevent uninteresting findings by design.	Use strong and non-obvious hypotheses; avoid win–lose setups; use strong baselines; compare more than two alternatives; be able to fail and/or generate surprises; use complete and representative conditions.
Provide evidence	Provide supporting data for all main conclusions.	Make chains of evidence clear; provide descriptive statistics; avoid easy/common errors in inferential statistics; ensure conclusion validity; report manipulation checks; use multiple, rich measures.
Narrate results	Explain results by anticipating and answering readers' questions.	Describe participants' interaction; speculate and provide data about "whys"; compare with known mechanisms and effects; give implications for researchers and practitioners; tie to hypotheses if possible; justify key decisions.
Bring an open mind to analysis	Explore alternative hypotheses and theories to understand data.	Explore alternative hypotheses; work against confirmation bias; discuss multiple interpretations of data.
Recognize limitations	Acknowledge and discuss limitations of setup, data collection, and analysis.	Discuss limitations; explain what could have been done differently (and how); discuss future research.
Respect participants	Treat participants, their time, and the data they create (behavior, comments, etc.) with respect.	Be ethical; don't waste people's time; aim for experimental realism; motivate participants; give a debriefing; allow participants to opt out at all times.
Be pragmatic	Any experiment is limited in its ability to say anything substantive.	Have a fallback plan; do not attempt all in one experiment; borrow and imitate excellent experiments; be creative in operationalizing variables; manage variability in performance; do pilot studies; share methods and results.

4

Designing Experiments

The design of experiments refers to the selection of the key components of an experiment (see Figure 1.1), and their organization into an experimental situation that participants experience and act in. Next we go through those components in turn.

4.1 Hypotheses and Theory

The role of hypothesis and theory in designing and running experiments is controversial. Here we discuss research hypotheses, which conjecture a relation between two or more variables. Some writers find it important even crucial to form opinions about the outcome of an experiment before running it. These opinions may be more or less detailed, but should be testable and justified, for instance by appeal to earlier work, relevant theories, or predictive models. Other writers find hypotheses less useful. They maintain that though a final written up version of an experiment may well include hypotheses, some research questions are exploratory and that an emphasis on hypotheses may bias our thinking. Sometimes, these two views are contrasted as "testing theory" and "hunting phenomena" [42], or as experiments that require theory by definition because they test it and experiments that may or may not

involve theory [142]. Deciding between these two forms is difficult, and requires that an experimenter compare the benefits and drawbacks discussed next.

We see a number of benefits of hypotheses. First, they help gain clarity about what one is doing and may help focus a research question. To do so, hypotheses must be (a) testable, (b) concise, and (c) name key constructs. For example, Nass et al. [104] hypothesized that "subjects will perceive a computer with dominant characteristics as being dominant" (p. 288). Gutwin and Greenberg [56] held the hypothesis that "better support for workspace awareness can improve the usability of these shared computational workspaces" (p. 511). The first example is testable because one may compare computers with and without dominant characteristics and expect a significant difference in participants' perception of dominance. That example names the key construct dominant, both as something that may be manipulated in computer interfaces (an independent variable) and as something that participants perceive (a dependent variable, assessed for instance by a questionnaire).

Second, formulating hypotheses helps a researcher think through what earlier work says about the experiment being designed. Thereby, hypotheses help summarize earlier work and use that work in motivating and designing the experiment. One way of doing so would be to use earlier theories to predict likely changes in dependent variables; another would be to use empirical findings to motivate hypotheses. In a paper on why people find it annoying to overhear conversations on mobile phones, Monk et al. [99] separated three explanations of the annoyance. They thought that people may be annoyed because (a) of louder noises, for instance, by ring tones and people speaking louder than in face-to-face conversations, (b) mobile phone conversations are more recently invented than face-to-face conversations, or (c) only half of the conversation is heard. Each explanation was motivated by earlier work and later discussed in light of the experimental data.

Third, hypotheses help report an experiment. An analogy may be made to the classic argument by Parnas and Clements [115] on how and why to fake a rational software design process. Parnas and Clements argued that even though we cannot do software development

by rational deduction of software design from requirements, we might benefit from faking the process, that is, from doing the documentation that would have been made in a rational process. Analogously, hypothesis may be useful for presenting experiments *as if* they had driven an experiment. For a reader of the report on an experiment such hypotheses provide structure and make it conform to the usual way of reporting experiments. Note, however, that if hypotheses are created after the fact one should not write a study up to generate the contrary impression [83]. That would make it impossible for readers to distinguish planned comparisons from accidental findings. For clarity of thought, we want to avoid this because it conflates difficult and potentially daring predictions from ad-hoc findings; for proper statistics, we want to avoid this because it increases the chance of reporting spurious findings as significant.

Fourth, hypotheses are tied to theory. Platt [117] presented what he called the Question, which any experimenter should be able to answer. It asks "But sir, what theory does your data disconfirm" (p. 352). The best time for a researcher to think though the Question is of course prior to running an experiment. In HCI, many theories may inform generation of hypotheses. Sears and Shneiderman [125] created predictive models of split menu performance prior to running a formal experiment. Nass et al. [104] used an extensive theory of how people communicate to motivate their study and derive several hypotheses. This is related to the earlier discussion of the "no surprises" objection to experiments.

Unfortunately, formulating hypotheses has several drawbacks. In psychology, Greenwald [52] argued that too often the null hypothesis of an experiment is formulated only to be disconfirmed. He suggested to design experiments where failing to reject the null hypothesis would also be a valuable outcome. Another way around this issue is to formulate alternative hypotheses, each supported by particular earlier findings or theory [117]. Thus, the experiment becomes an attempt to disconfirm one or the other equally plausible hypotheses. These concerns illustrate the need to be wrong occasionally and to learn something from being wrong. A related, older idea is that of Chamberlain [22], who spoke about multiple working hypotheses. His idea

was for experimenters to come up with multiple hypotheses to avoid becoming too attached to one hypothesis and thereby run the risk of overlooking evidence against that hypothesis. The recommendation for HCI experiments is to do strong comparisons that may fail, and where failure will teach us something valuable (see Table 3.1). As mentioned earlier, I believe too many experiments in HCI cannot fail or — equivalently — that their results are given a priori. Across different scientific disciplines, an increasing number of papers confirm their hypotheses [36]. Thus, negative results (i.e., failures to find differences or replicate earlier findings) seem to be less and less frequently published. For authors, this trend might inspire more daring studies; for reviewers and editors, it might inspire more lenience for papers with negative results, if these are designed as strong comparisons.

A strong argument against the reliance on theory is that it may blind us to interesting findings. Greenwald et al. [52] argued that researchers who depend heavily on theory are more likely to revise experiments (e.g., change procedures, predictions, or analysis of data) if finding disconfirming evidence rather than to revise theory. They presented examples of how such confirmation bias obstruct progress in research. Garst and colleagues [43] presented data that participants who were given a hypothesis about differently sized letters placed in various positions on a card generated fewer, simpler alternative hypotheses than participants who were not given any hypothesis. Holding a hypothesis may blind us to interesting findings; the recommendation for analysis from Table 3.1 is to bring an open mind to analysis or use multiple theories to analyze data.

In sum, the view here is that not all experiments need hypotheses or theories (and in some cases it is not feasible to create hypotheses or impossible to find theories to draw on). But all experiments may benefit from thinking through whether hypotheses may be formed and which theories or models that can help us think about the relationship between the independent variable and the dependent variable before actually running an experiment. In my opinion this is done too rarely, in particular for experiments where much control is exercised over the experimental situation and where studying the context of use or richness of behavior are not contributions.

4.2 Independent Variables

Given a research question and some hypotheses for an experiment, it might seem easy to design the conditions that participants will experience, that is, the levels of the independent variables. It is not. One concern in choosing levels of the independent variables (or conditions) is to ensure that they match the key ideas (or constructs) of the research question. The process that determines the specific make up of independent variables is called operationalization, and the extent to which that process creates conditions that reflect well the constructs of the hypotheses and research questions is called construct validity (e.g., [127]). Here conditions are about the substantive domain or content (e.g., an actual user interface), and constructs are about the conceptual domain or ideas (e.g., an interaction paradigm); McGrath [96] further elaborates this distinction. Figure 4.1 illustrates this process and the four questions that we discuss next.

The conditions must be *complete and representative instances* of the constructs being studied. Say a group of researchers wants to understand the relative effectiveness of four techniques for animating transitions (as did Dragicevic et al. [32]). If they study slow-in/slow-out animation, the operationalization of that technique should include all essential characteristics of slow-in/slow-out animation. Dragicevic

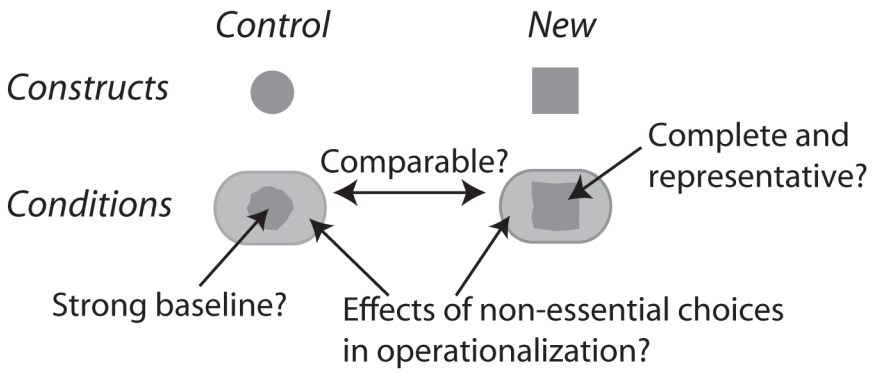

Fig. 4.1 Four questions to guide finding good independent variables. Operationalization is about turning constructs (e.g., a control and a new user interface, represented as a circle and a square) into conditions (shown as rounded rectangles that contain reflections of constructs in addition to other issues).

et al. [32] argued this representativeness by explaining how many commercial and research interfaces use slow-in/slow-out animation and presented an operationalization supposed to be typical of those interfaces. Experimenters may use focused research questions to argue completeness; earlier work may help argue why a particular condition is representative of the more general construct.

Unfortunately, most operationalizations of a construct are incomplete or biased. Shadish et al. [127] suggested that so-called mono-operation bias poses a serious threat to construct validity. Mono-operation bias occurs whenever an experiment uses just one operationalization of a construct. Such a single operationalization may underrepresent the theoretical construct as well as introduce factors in the experiment that are not central to the construct. The recommendation here is to study more than two levels of an independent variable: using alternative operationalizations of the construct may increase the validity of our concluding at the general, construct level. Also, multiple operationalizations may help move beyond the win–lose setup of many experiments. A practical argument for using several levels is that even if a level of the independent variable (say, a new interface paradigm) does not lead to any difference (say, over a control interface), it is still possible as a *fallback plan* to generate interesting data about the differences between the instances of the construct. In a sense this advice follows from the earlier discussion of Greenwald et al. [52] and of Chamberlain [22]. As an example, Paek et al. [113] used three versions of an experimental information retrieval system (and a unmodified baseline) to identify the dynamic layout technique that worked the best for presenting web search results; many experimenters in their place would have chosen only to study one version and the baseline.

Experimenters often hypothesize that UI constructs improve state-of-the-art or older UI constructs, acting as control or baseline. The aim is to be complete and representative, but also to ensure *strong baselines*. Munzner [103] discussed what she termed "Straw Man Comparisons", cases where authors compared their interfaces against outdated work, rather than the state-of-the-art approaches. In comparing graph layout algorithms, for instance, authors might compare against naïve spring algorithms, rather than newer and better approaches. Though Munzner

wrote about information visualization, her notion applies to HCI in general. In studies on usability evaluation methods, for instance, Hornbæk [68] argued that many studies that compare usability evaluation methods employ a win–lose setup, where a novel method is compared with a dated or inferior baseline. Landauer [86] argued that comparisons with prior computer and non-computer methods are of "paramount importance" (p. 209). The work on SuperBook [35], an experimental hypertext browser, provides an excellent example of comparisons with strong baselines. In particular, SuperBook was compared with a print version that contained the same textual information.

In turning a construct into actual conditions, researchers often want to add UI features or functionality that are not relevant to the research question, but useful or necessary for the tasks or activities to be done. The question here is about *whether non-essential choices in operationalization bias or otherwise affect the main comparison of interest.* A variant of this problem specific to HCI concerns the difference between tools and techniques, or between what may be called "whole package" and "essential features". One way to proceed with experiments is to compare entire tools for working in a particular domain or supporting a particular activity. An interface provided to participants would contain both features that are being varied (say, a search presentation technique) and features that are not varied but expected in real interfaces (say, whether or not search requests are being spell-checked). If the latter features were included, participants' experience would be close to what would be provided in realistic use. The "whole package" approach was used by Egan et al. [35] in their work on SuperBook, which combined search, table of contents, and annotations. Another way to proceed with experiments is to isolate the features of interest and ignore or leave out other types of feature that would be expected in a realistic tool (as done in the paper by Dragicevic et al. mentioned earlier [32]). The strength of this way is to isolate the main variables of interest and to minimize variation due to other features of the tool. Similar considerations apply to experiments whose independent variables are not interfaces.

Figure 4.1 suggests another key question, namely whether levels of the independent variable (say, two versions of an interface) are *similar*

(or comparable) in all respects except those one is interested in manipulating and concluding about. If they are dissimilar (or confounded), any effects on the dependent variables cannot be attributed to what we would like them to be attributed to, namely levels of the independent variable. Whereas this question concerns the operationalization of constructs, some writers (e.g., [127]) use the term internal validity rather than construct validity. In experiments on HCI that compare interfaces, typical questions about operationalization of independent variables include (assuming, of course, that these concerns are not part of the research questions or hypotheses):

- Is the functionality equivalent?
- Is the setting similar?
- Is comparable information available in the interfaces?
- Are the size and structure of the interfaces similar?
- Are comparable hardware used (e.g., for input and output)?
- Is training with the interfaces comparable in complexity and duration?
- Are the instructions for operating the interfaces similar in scope and complexity?
- Are the allotted time and criteria for success similar across interfaces?
- Is the starting point identical for participants who receive different treatments?

The literature contains many examples of choices of conditions that were unsuccessful, in the sense of leading to comparisons that did not allow for the conclusions intended. Lam and Munzner [85] gave examples from information visualization, where conditions being experimentally compared differed in their support for displaying the same amount of data, in how information was visualized, in their ability to show comparable level of details, and in having markedly different complexity of interaction. In all cases, the authors did not intend to study these differences, but introduced them by insufficient care in their experimental design. Gray and Salzman [49] showed how comparisons of evaluation methods were deeply flawed, as judged from a catalogue of concerns about validity. For instance, one problem pointed out by Gray and Salz-

man [49] was that some studies of usability evaluation methods had specific groups of users apply a particular method. Usability experts would do a heuristic evaluation and their performance be compared with undergraduates who did a keystroke-level model (KLM) evaluation. With such a design, effects of evaluation method and evaluator background are confounded, rendering strong conclusions impossible. In another instance, researchers had allowed evaluators much longer time to apply one method compared to the time allotted to evaluators using another method. In that case, strong conclusions about differences in evaluation methods were impossible, because they were confounded with time. Other research questions or domains typically raise additional questions about comparability.

Unfortunately, making conditions comparable is hard. In one study, we were comparing interfaces for reading text on computers [69]. One such interface offered both an overview of the entire document and a detailed view of the text (a so-called overview+detail interface). Another interface offered a baseline approach to presenting the text (a so-called linear interface), similar to how text is presented in most word processors, document readers, and web browsers. It was hard to make the overview+detail interface differ from the linear interface only with respect to the overview. On the one hand, the layout of the text — in particular the line length — should be similar, because layout greatly influences reading time and reading processes. On the other hand, the screen real estate should be similar, because the comparison otherwise might just show that more space is better. This might lead to the "no surprises" objection mentioned earlier [37].

Another recommendation is to do *manipulation checks*. A manipulation check attempts to gauge whether the independent variable works as intended. For instance, one may test if participants understand and use a new UI feature as intended by the experimenter. Jeffries et al. [76] described what turned out to be an important manipulation check. They were studying usability evaluation methods and asked participants to report how they identified usability problems. Surprisingly, participants reported finding problems both by prior experience and as "side effects" of using a technique (e.g., noticing one type of problem while following the method to look for another type of problem).

Later papers have used this observation to argue that what is going on in the use of usability evaluation methods is not only (or mainly) method application (e.g., [68]). Even though Jeffries et al. did not name this a manipulation check, their data suggested that the manipulation of the independent variable (evaluation method) was not working as expected. Recently, my colleagues and I [40] investigated perceptions of usability across cultures. We sampled Danish and Chinese respondents and claimed they were representative of two different cultures. Although we knew their national background, upbringing, and so on, we did not test this difference using one of the many available questionnaires on cultural differences. Having done so would provide a clear manipulation check of culture. In sum, manipulation checks give a post-facto way of ascertaining and understanding how independent variables work in the experimental situation, in a sense *providing evidence* that the manipulation works.

In many experiments there is a desire to manipulate more than one independent variable. We deal with one such independent variable, task, in a subsequent section. Here just a few remarks on multi-factor experiments. Many interesting phenomena in HCI depend on interactions among variables, as already discussed by [86]. For instance, task complexity may interact with interface, meaning that one interface works well for complex tasks whereas another works well for simple tasks. To detect such a pattern, one needs multiple factors in a single experiment (or several experiments). Unfortunately, including multiple factors in an experiment comes at an expense. They complicate the structuring of experiments as well as their analysis. The latter point, in particular, makes presenting results in a manner that is clear and understandable hard. Thus, my thinking is mostly to *simplify the design*, reducing the number of independent variables to two or three, and investigate other variables of interest in separate experiments.

Finally, selecting and operationalizing independent variables is also about *being creative*. Many important experiments in HCI are based on ingenious ideas or operationalizations of independent variables. Gould et al. [48] wanted to study real-time use of speech-processing software, but no such software existed. Thus, they faked the software by having a person listen to the input to the software and transcribe it (using

a Wizard of Oz approach). Cockburn and McKenzie [25] compared 2D and 3D user interfaces for web pages and created a physical 3D interface using photos of webpages that were suspended from a system of fishing-lines. Monk et al. [99] wanted to understand reactions to mobile phone use, in particular the reaction to being able to hear only one part of a conversation. They had actors play out different scripts to be able to combine the field setting with relatively uniform situations to participants.

4.3 Structuring Experiments

Structuring experiments implies designing how participants will be subjected to the levels of the independent variables and the tasks. The aim is to ensure internal validity, that is, the ability of an experiment to attribute differences observed in the dependent variables to manipulations of independent variables [127]. Another way to think about internal validity is to consider whether any differences in the dependent variable would be present without variation in the independent variable. A frequent example of such a validity concern is maturation [127], that is, changes in participants' performance over time. A study might be invalid in this sense by presenting participants with the control interface first and a novel interface second. We might see a difference between interfaces even without the novel interface because participants mature, for instance, by gaining experience. Put another way, the main concern in structuring experiments is to rule out by design alternative hypotheses to the one you are testing.

A key tenet of classic thinking about experiments is the use of *randomization* (e.g., [42]). The idea is to assign participants at random to levels of the independent variables. That way, factors that are not of interest to the experimenter are evened out between conditions, improving the experiment's internal validity. In practice, randomization is often done by rolling dice, consulting a list of random numbers, or using a random number function. This is to avoid the many misfortunes described in the literature where what seemed to be random procedures for assignment was found to show systematic biases. Randomization has the big advantage that it is easy to do and that it deals also with

factors that influence the experiment, but are unknown. An alternative to randomization is *blocking*. The idea here is divide the experiment into blocks that are similar, for instance so that a block contains two versions of a user interface. We may then repeat blocks and get a more even spread of the variation due to user interfaces compared to a non-blocking design. Cockburn and colleagues [24] used both these approaches to structure an experiment for investigating menu navigation. They were interested in modeling menu navigation but needed to estimate a set of parameters for the model. Among other things they varied the type of menu (three levels) and the length of menus (four levels). The experiment used all three types of menu, in an order determined at random. That way, the levels of the menus are likely to be evenly distributed among the first part and the last part of the experiment. The length of menus was blocked, so that participants did three or seven blocks of menu selections, where each block contained selections of all items in the menu (determined by its length). That way, Cockburn and colleagues ensured that selections in the various parts of the menu were evenly spread over the course of the experiment.

A key distinction in assigning participants to interfaces, task, and levels of other independent variables is between between-subjects and within-subjects designs. In a *between-subjects design*, each participant experience only one level of the relevant variable, say, one of the interfaces being investigated; in a *within-subject design*, each subject experiences all levels of the variable. Both are used in HCI, but within-subject experiments appear to be the most frequent. For instance, in a sample of 73 studies, 60% used a within-subjects design and 32% a between-subjects design; the remainder used a mixed approach [70].

Between-subject experiments offer benefits. First, they are simple to describe, simple for participants, and simple to analyze statistically. Second, they are the only option for some experiments. For instance, many learning effects can only be studied between-subjects and some experiments employing deception cannot be done within-subjects. The study by Nass et al. [104] on the perception of computer personalities mentioned earlier used a between-subjects design. The reason for this design was likely that (a) analysis would be simplified, (b) participants would not become aware of the purpose of the experiment, which they

might if they had seen two interfaces that differed only in style of communication, and (c) that only one task needed to be developed because each participant only used one interface to do one task. Had a within-subjects design been used, Nass et al. would have had to construct several tasks that were equivalent, yet not so similar that having done one would change how participants did the next.

Within-subject designs also offer several benefits. First, they offer statistical advantages. Because participants are subjected to more than one level of the relevant variable, the variability of the estimates of that variable will typically be more precise (though not always, see Keren [82]). More precise estimates lead to higher power, the ability of an experiment to detect differences between conditions. Second, within-subject experiments offer practical advantages. Because participants can use several interfaces (or whatever is being manipulated), fewer participants are needed for the experiment (but sessions may end up being longer). Within-subject designs face several methodological issues; Greenwald [51] summarized three of particular importance. Practice is an issue because participants may learn from being exposed to several instances of an interface or task and thereby gain proficiency. In some experiments, practice is of importance; in others, it may be undesirable. Greenwald also discussed sensitization, which concerns the possibility of participants figuring out the purpose of an experiment. When seeing variations of the treatments, it is much easier for participants to form opinions about the differences among treatments. Finally, carry-over effects are about the influence one treatment may have on subsequent treatments that a participant experience. For instance, one may see asymmetrical transfer between conditions, meaning that user interface 1 influences user interface 2 more than user interface 2 influences user interface 1.

Several tactics and standard practices help experimenters stay clear of some concerns about internal validity. The act of counterbalancing helps remove many potential confounds resulting from the ordering of the levels of independent variables. Latin squares and Greco–Latin squares offer general solutions for distributing the effect of order and combination of two or more independent variables. The number of papers that have been rejected because one or more of these tactics

Fig. 4.2 Counterbalancing with Latin and Greco-Latin Squares. Panel A shows a within-subjects design for three user interfaces (ui1–ui3), each used in a sequence of sessions (session 1–session 3) by participants (user 1–user 3). Panel B shows a Greco–Latin Square that crosses user interfaces with tasks (t1–t3). Panel C shows a $4 * 4$ Latin square, for a situation with four user interfaces (or four other levels of an independent variable). That square is balanced for first-order effects in that each user interface is followed in the next session by any other user interface the same number of times.

were not followed is large. Figure 4.2 shows some examples of Latin and Greco–Latin Squares. Most statistics programs can produce such squares (e.g., SPSS, R).

4.4 Participants

Arguably the most important question in selecting participants for an experiment is *who should participate*. Hornbæk and Law [70] found that in a sample of 73 HCI experiments, half were conducted with participants that had much experience with the task to be performed and about a third were done with participants with no experience. The key issue here is whether or not participants bring expertise, insights, aspirations, expectations, or specific work competencies to the experiment that is important. Campbell [18] expressed this in a negative manner by asking whether "the specific work experiences of the subjects influence the phenomena being studied in such a way that they confound the results of the study" (p. 276). Positively put, selecting participants is about finding people with the characteristics necessary to address the research question. Key factors to think through include domain experience, IT experience (e.g., in general or with an interface being tested), personal characteristics (e.g., gender), abilities (e.g., in thinking, perception), attitudes, and motivation (e.g., financial, intrinsic).

The question of whether students may validly be participants in experiments recurs. Barkhuus and Rode [10] found that about half of a sample of studies from the ACM CHI conference used students as participants; Sjøberg et al. [131] found that 81% of a sample of 113 articles on software engineering used students. On the one hand it is well known that college students differ from other potential groups of participants: They have stronger cognitive skills, have developed less strong attitudes, and are more likely to follow authority [126]; most are also shaped by living in western, industrialized, and democratic societies [62]. On the other hand, in the context of organizational psychology Campbell [18] lamented that "In spite of the gaps in the research record mentioned earlier, the message is clear: the data do not support the belief that lab studies produce different results than field studies. Perhaps college students really are people" (p. 276). Thus, in itself having students participate in an experiment may not matter to a study; conversely, using experts in a particular domain may not matter for some tasks. One may think the same way about several other systematic methods for choosing participants. Volunteer participants, for instance, show many differences to the general population in tending to be more well-educated, more intelligent, of higher social class, more arousal seeking, and more approval motivated [120]. The differences between the general population and those who are willing to complete tasks for micro payments (e.g., on Amazon Mechanical Turk) are just beginning to be understood (e.g., [114]). In all cases, what matters is to think through what characteristics of participants that matter.

The selection of participants is often guided by a wish to generalize any effects of the independent variable to other persons; Shadish et al. [127] referred to this as an instance of external validity. The key question is to understand if characteristics of participants may influence the size or direction of effects, when one is considering broader or more focused groups of users. When designing experiments, one should think about whether the selection of participants affects generalization: Will conclusions hold also for the general population, the average user, the expert user, prospective users, or any other group that the experimenter would like to conclude about?

An often-pondered question is *how many participants to use*. A couple of answers may be given. The technical answer is that power analysis [26, 27] may be used to estimate the probability that one detects a difference in a dependent variable between levels of the independent variable if one knows (or can qualify a guess about) the magnitude of the effect one is looking for. Power analyses are depressing reading. Typically many participants are required to achieve a reasonable power (say, an 80% probability of finding a difference). To detect medium-sized differences between two conditions at this probability, one would need 64 participants in each condition in a between-subjects experiment. Medium-sized effects found in the HCI literature include differences between broad and deep menus or between selecting with mouse and keyboards.[1] Fortunately, many studies can obtain repeated measures and use within-subjects designs, both of which reduce the number of participants needed. In the example above we may obtain 10 repeated measures in which case one would need only 25 participants per group (assuming that observations are correlated at $r \approx 0.3$, the case for instance in Jakobsen and Hornbæk [73]). With that number of repeated measures, a within-subjects experiment would require about 20 participants in total. Most statistics packages can assist with power analysis, as can the free tool G*Power (Faul et al., 2009).

The pragmatic answer to the question about how many participants to use is that HCI studies typically employ 20 participants (medians from [10, 70]); substantial variations of course exist among experiments with different purpose and in different application areas. Incidentally, in a critical review of psychology experiments, Simmons et al. [130] also recommended 20 persons per condition because "samples smaller than 20 per cell are simply not powerful enough to detect most effects, and so there is usually no good reason to decide in advance to collect such a small number of observations." (p. 5). Note that the technical

[1] These effects differ between studies. For instance, Larson and Czerwinski [88] found a medium effect size (Cohen's $d = 0.47$) between their slowest (a deep menu) and fastest (a broad menu) conditions. Card et al. [20] studied pointing to a word on a page of text and found a medium-sized difference between mouse and step keys (a step key may be used to navigate a word, sentence, or paragraph, $d = 0.48$), and a medium to large difference between mouse and text keys (similar to arrow keys, $d = 0.70$).

and pragmatic answers above are unrelated to recommendations on how many users to use in a usability test (e.g., [71]). These recommendations are about formative evaluation, not about the summative evaluations that are the focus of this monograph.

4.5 Tasks and Activities

Most experiments in HCI have participants engage in some form of structured activity, typically referred to as tasks. A key characteristic of experiments is that these tasks are not initiated or motivated by participants themselves, but take place because of the experiment [42, 96]. Some studies that concern open-ended or non-work-related activity often refrain from using the word task; these studies still have to create a meaningful experimental situation and to instruct participants what to do, meaning that they face many of the same concerns as studies that use tasks. Our use of the word task below also covers such cases. Note that many experiments include several tasks and treat these as an additional independent variable when structuring experiments and when reporting them. Many of the concerns listed in the previous section (e.g., completeness, representativeness, and non-essential features) also apply to thinking about tasks and activities.

One may select tasks in many ways. One is to select tasks that are *representative of what users would do outside of the experiment*. Munzner [103] discussed selection of tasks in information visualization and wrote "A study is not very interesting if it shows a nice result for a task that nobody will ever actually do, or a task much less common or important than some other task. You need to convince the reader that your tasks are a reasonable abstraction of the real-world tasks done by your target users" (p. 147). One way of ensuring representativeness is to use tasks that users have been observed doing. Meister [98] recommended that evaluation proceed from domain studies of actual task attempts. Another way of ensuring representativeness is to proceed from taxonomies of tasks or other syntheses of tasks that users do or want to do. A third way is to only prescribe activities loosely and then later on figure out what was actually done (e.g., [111]). For instance, one may leave parts of the task to be filled in by partici-

pants: This gives more realistic and interesting tasks, at the cost of higher variability. Gutwin and Greenberg [56] selected tasks for evaluating awareness-support in a groupware system, aiming to find representative tasks. They wrote: "Tasks were designed to mimic episodes and activities that we observed in face-to-face collaboration" and that "we wanted realistic tasks that were likely to occur in a wide variety of workspace activities" (p. 512). From these desiderata Gutwin and Greenberg developed and used three tasks in their experiment.

Another approach to selecting tasks is to *use simple tasks* that capture the essence of what is being investigated. The idea is to reduce variation and remove non-essential features of a task; this idea is similar to the approach for selecting independent variables that was earlier referred to as essential features. For instance, many studies of pointing techniques use the ISO multidirectional tapping task [133]. This task requires participants to tap circular or square targets arranged in a circle. It does not represent pointing in the wild, but is widely accepted as a useful task for experiments. Dragicevic et al. [32, p. 2012] explained the reasoning in adopting a simple task for a study of animation as follows:

> This is an elementary low-level task, ensuring that if users are unable to perform it, then more complex tasks — e.g., following multiple independent objects or groups of objects — will be equally or more difficult. It is also safe to assume that many higher-level tasks will be difficult to perform if single objects cannot be tracked.

A third approach uses *task-specific hypotheses*. Tasks in HCI often lead to significant differences in performance, outmatching even individual differences. Therefore, this approach selects tasks for which experimenters have specific expectations or hypotheses. For instance, Alonso et al. [5] created 31 tasks for a comparison of Lifelines — a visualization of time-related medical data — to a control interface. For each task, Alonso et al. presented a hypothesis for or against Lifelines or the control interface. Task-specific hypotheses is related to the idea

of boundary variable discussed by Fromkin and Streufert [42]. Their idea was that experimenters can identify factors that can change a relation between independent and dependent variables, giving a richer understanding of the phenomena under study compared with merely investigating if a particular relation exists. Tasks may be such a variable and task-specific hypotheses may help characterize the boundaries of a specific effect.

Deciding between these approaches is hard. On the one hand, representative tasks may contain irrelevant elements and may lead to greater variation in performance; hypotheses may also concern only a part of such tasks. On the other hand, using simple tasks requires an argumentation for why they are relevant for real work (for instance like the one by Dragicevic et al. quoted above). One critique of experiments in HCI is that they use too simple tasks. Among a sample of 73 empirical studies in HCI, Hornbæk and Law [70] showed how tasks were mainly low complexity (say, perceptual or motor tasks) and only infrequently of high complexity, for instance requiring participants to problem-solve (19% of the studies). Whereas some research questions are inherently about low complexity tasks, I think that probably more studies could use representative tasks. Using task-specific hypotheses seems the least frequent approach of the three discussed here, but it is powerful for scoping known effects and for understanding better why and when something happens (because variation in tasks can provide data about why and when).

When selecting tasks, there is ample opportunity to *build on the works of others*. Whittaker et al. [140] argued that HCI overemphasize radical invention, that is, novel and bleeding-edge technologies for rare or somewhat contrived problems. They suggested that this emphasis lead to a lack of comparability across studies, and to studies that re-invent tasks. Thereby, "researchers can end up proposing radical solutions to things that users do not consider major problems and can neglect major problems that users do experience" (p. 79). Whittaker et al. proposed a reference task agenda for HCI. Although such an agenda is infrequently pursued on a large scale, standardized tasks or taxonomies of task are available in information retrieval (e.g., text retrieval conference, TREC), information visualization (e.g., Visual

Analytics Science and Technology Challenges), input studies (e.g., [133]), and many more areas. Yet, the tasks people do change over time, and often technologies and tasks are intertwined (so that new tasks emerge with new technologies). So it is not realistic to expect reference tasks available for all experiments.

Finally, participants' *understanding of tasks and their involvement in the experimental situation are important to all studies.* If participants do not understand the task, their performance is likely to vary a lot and to be of little value to the experimenter. Some researchers maintain that understanding tasks implies being aware of the criteria for evaluating performance. In many studies on pointing, for instance, this has lead to researchers to instruct participants to "work as fast as possible while still maintaining high accuracy" [31, p. 217]. Others maintain that experimental realism — the extent to which participants experience the experiment as meaningful and involving — is key to useful insights. Experimenters can try to achieve the above by pilot testing, careful debriefings that elicit participants' view of the experiments, clear instructions, and intrinsically motivating tasks.

4.6 Setting

The setting in which experiments take place requires some comments; in part because setting is an important issue to experimenters, in part because setting is much debated both in HCI (e.g., [84]) and outside (e.g., [6, 45]). One consideration for setting is whether experiments take place in the lab or in the field. In lab experiments, the setting is controlled and the effect of external influences minimized. In field experiments, the setting is real, although the experimental manipulations are instigated by the experimenter. The view taken here is that neither choice of setting is better than the other; rather, they have relative benefits and drawbacks.

On the one hand, the lab setting offers great potential for restricting the influence of extraneous variables and for collecting fine-grained data. It is likely to reduce variability in performance and hence to increase power. On the other hand, the field setting is attractive. It allows us to discover phenomena that were not anticipated and to study

activities too complex to bring into the laboratory. Oulasvirta [111] surveyed the practical challenges of field experiments well, including how to gather better data, test predictions in the field (rather than just hunt phenomena), and how to study the effects of the environment on interaction.

Sometimes a lab setting may be relaxed. Meister [98] suggested that experiments could "introduce into the experiment, to the extent possible, conditions that are representative of the operational environment"; Landauer [86] similarly talked about robustness over variation, where the setting provides one important source of variation. The attention to setting sometimes turns into an obsession with mundane realism, the physical similarity to the real setting, for instance with usability labs and experimental settings constructed to resemble living rooms. However, as argued by Berkowitz and Donnerstein [12] experimental realism matters more than mundane realism.

The discussion of setting is sometimes confounded with the question of method; some arguments against lab experiments slip into arguments against experiments. My view is that these are separate issues, as evidenced by many excellent field experiments in the literature (e.g., [64, 99]).

4.7 Dealing with Other Factors

In the case of extraneous variables (called *other factors* in Figure 1.1) the strategies mentioned earlier (randomization, blocking) may also be used. In addition, we may *control* factors, meaning that we require a set level of expertise or allow only a certain gender to participate or conduct all our experiments in the same room. *Matching* is similar to blocking in its approach to dealing with other factors. It means that the experimenter matches participants in different conditions on some factor. For instance, participants may be matched on gender so that whenever a female is assigned to one condition, a female is also assigned to the other condition or whenever participants with high mental rotation skills are assigned to one interface, a participant with low such skills are assigned to the other interface. It is also possible to measure factors that are of interest and subsequently use those measures to *statistically*

adjust the effect. This is done by removing the effects of variation in a variable statistically after running the experiment [96]. The idea is to obtain some measure of the factors (say, dexterity, visualization ability, or screen resolution) and include them in statistical analysis, using for instance analysis of covariation. Burnett et al. investigated gender differences in programming. In one experiment they had participants with different levels of confidence in themselves (as measured by a validated self-efficacy scale) use a new spreadsheet environment. Burnett et al. used self-efficacy as a covariate in their analysis, allowing them to conclude about the spreadsheet environment in ways not influenced by initial self-efficacy scores.

4.8 Choosing Dependent Variables

The choice of dependent variables follows from the hypotheses of an experiment. It will also, however, be shaped by the application domain, interface technology, context of use, related work, and so forth. The main concern in selecting dependent variables is about construct validity. Construct validity is in part about the extent to which the actual measures collected reflect what the researcher intends to measure, or about "making inferences from sampling particulars of a study to the higher-order constructs they represent" [127, p. 65]. Thus, all obtained measures or scores are typically involved in an act of reasoning where they are taken as indicators of a general or theoretical concept: the extent to which this act is justified concerns construct validity.

One may separate the two issues of *conceptualization* and *operationalization* [4, 127]. On the one hand we need to understand the constructs of interest. For instance, while learnability of a user interface is (superficially) an easy-to-understand quality, defining it is much harder. Grossman et al. [54] showed how the literature displays many different understandings of learnability. If an experiment does not clearly conceptualize learnability the validity of any inferences from that experiment may be reduced because learnability may mean many different things. Similarly, task completion time is very easy to measure, but it may not be the best conceptualizations of the qualities of an interface that an experimenter seeks to establish. Studies vary in whether they see low

task completion times as good (minimizing resource expenditure) or bad (expressing a lack of engagement), see Ref. [67].

On the other hand we need to develop ways of actually measuring outcomes of an experiment, operationalization. This is very hard. Hornbæk [67] showed how researchers in HCI have devised a myriad of questionnaires for measuring subjective satisfaction, many of which have been shown unreliable or otherwise problematic in their operationalization [70]. Construct validity thus requires an experimenter to carefully think over and define key constructs in addition to reasoning about whether actual measures and measurement procedures are representative of those key constructs.

Measures of usability will form the dependent variables of many experiments, in particularly those comparing user interfaces. One prevalent way of understanding usability is to see it as quality-in-use of an interactive system, that is, "the user's view of the quality of a system containing software, and is measured in terms of the result of using the software, rather than properties of the software itself" [13, p. 92]. Whereas dependent variables may be named differently (e.g., workload, performance, user experience), the umbrella term usability is used below.

The research on usability raises important questions for experimenters. It has been shown that *usability is multi-dimensional*. Early models distinguished five groups of measure [108, 128]: learnability/time to learn, efficiency/speed of performance, memorability/retention over time, errors, and subjective satisfaction. ISO 9241-11 [72] distinguished three types of measure: effectiveness (e.g., accuracy), efficiency (e.g., time), and satisfaction. A related insight is that that objective and subjective indicators of usability may differ empirically and conceptually (e.g., [70, 124]). For instance, Hornbæk and Law [70] found that users' perception of outcomes did not correlate with actual measured outcomes; Hassenzahl [59, p. 9] argued that subjective and objective qualities are fundamentally different:

> Experience is subjective. It emerges through situations, objects, people, their interrelationships, and their relationship to the experientor, but it is created and remains in her or his head. Given that, it may be not

> matter how a product is objectively, its quality must also be experienced to have impact.

One way to use these results is to assume that dimensions of the usability construct are relatively independent and to collect measures on all dimensions (see [41] for illustrations of this assumption at work).

Selection of usability measures may be inspired by *catalogues of usability measures* [9, 44, 67, 137]. Table 4.1 contains a sample of often-used measures. The main point here is to help experimenters reason about potential measures, rather than to provide a cookbook for running experiments. Many of the instruments for collecting such measures have been carefully developed and validated, a sure and important way of using earlier work. Jarvenpaa et al. [75] suggested "in the short run, modify and use, as much as possible, previously used and validated instruments; develop your own only if absolutely necessary" (p. 152).

While such catalogues give inspiration, a few additional considerations should be mentioned. First, studies in HCI could use richer and more complex measures of outcome. Hornbæk [67] argued that measures at the macro-level are too infrequently used in evaluations of interaction with user interfaces; this was echoed in the earlier discussion of complex tasks. Macro-level measures span hours or months, are cognitively and socially complex, and are typically about effectiveness or satisfaction (rather than efficiency). Complex measures were used in one of the studies of SuperBook [35], where participants wrote essays about features of the statistics system that SuperBook described. These essays were subsequently graded by an expert in statistics and by tallying the number of facts mentioned both in the essay and on a master checklist. Second, composite dependent variables are in my view much harder to interpret than non-composite ones. A composite variable could be the F-measure used in information retrieval (which integrates measures of precision and recall) or quality normalized by time (typically obtained by dividing a measure of task quality measure with task completion time). Such measures ease calculations and analysis, but may bewilder readers and hide detail. Third, multiple measures of the same construct increase reliability and strengthen the validity of claims about constructs. Using just one operationalization of

Table 4.1. Typical dependent variables in experiments in HCI (based on Refs. [9, 44, 67, 137]).

Construct	Definition	Example
Accuracy	Errors in trying to complete a task (e.g., task completion) or in the task results (e.g., spatial accuracy).	Proportion of corrects trials when using a mouse to steer through a tunnel [2].
Completeness	Amount or magnitude achieved in task solution (e.g., on a secondary task).	How completely a design task was covered [110].
Outcome quality	Assessments of the quality of the outcome of interaction (e.g., by learning assessments, expert rating).	Expert grading of essays written by the use of SuperBook or a control interface [35]
Time	Time taken to complete parts or the whole of a task.	Time spent in various parts of a design task solved with and without a shared text editor [110].
Effort	Resources expended to complete a task (e.g., communication effort, steps taken).	Steps taken in navigating a hierarchy [88].
Learnability	Easy to learn to operate an interface (e.g., to a specific criterion or for intermittent use).	Henze et al. [63] evaluated improvements of touch-type keyboards and measured learnability as changes in error rate over time.
Preference	Users' preference among interfaces (e.g., as indicated by rank ordering, rating, or implicit preference).	The interface users chose for a final task, after they have gained experience with a range of interfaces [64].
Workload	Subjectively experienced effort (e.g., as reported in questionnaires) or objective indicators of workload (e.g., pupil dilation).	Pirhonen et al. [116] measured workload while participants walked and used a mobile device; NASA's TLX was used [58].
Satisfaction	Assessment of users' satisfaction with an interface (e.g., through QUIS [23] or CSUQ [91]).	Chin et al. [23] used QUIS to compare liked and disliked products, as well as menu and command-like interfaces.
Affect	Assessment of users' affect while using an interface (e.g., with the self-assessment mannequin, SAM [87]).	Mahlke and Thüring [94] studied the perception of portable audio players using SAM, along with other measures.
Appeal	Users' perception of beauty, appeal, and aesthetics in interfaces or interactions (e.g., measured by Visual Aesthetics of Website Inventory [101]).	Lavie and Tractinsky [89] used questionnaires to measure users' perception of classical (e.g., beauty) and expressive aesthetics (e.g., originality) in web pages.

(*Continued*)

Table 4.1. (*Continued*)

Construct	Definition	Example
Fun	Users' experience of enjoyment while using an interface.	Mueller et al. [102] used a questionnaire to evaluate bonding and fun in exertion-based interfaces.
Hedonic quality	The experience of non-task related quality, such as novelty and stimulation (e.g., as measured by the AttracDiff2 questionnaire [60]).	Hassenzahl and Monk [61] studied the relation between beauty, usability, and hedonic quality on web sites, using AttrackDiff2.

a construct faces a mono-method threat to validity [127]. It means that we are more prone to not measuring what we think we are measuring if we use just one indicator for a construct. Thus, whenever possible, use several operationalizations of key constructs. For instance, Olson et al. [110] described how they developed a quality measure of designs for an automated post service through extensive discussion among researchers and designers; a rating form was constructed based on three aspects of design. Fourth, a useful notion in thinking about dependent variables is critical parameters [105]. A critical parameter is a performance indicator that captures aspects of performance that are critical to success, domain/application specific, and stable over variations of interface. Part of the challenge in applying catalogues of measures is to ensure that at least some measures chosen are critical in the above sense (and not just generic time or error measures). Fifth and finally, the strict definition of experiment proposed earlier means that measures of usability will always be relative, say compared to another interface or a base level, never absolute. Thereby one avoids the temptation to infuse meaning into an absolute usability score such as the average of numeric answers to a usability questionnaire. In conclusion, we recommend using *multiple, rich dependent measures*. Preferably, they should also have straightforward interpretations as to their relation to quality-in-use.

While we have mostly discussed the validity of usability measures, a word about reliability is also necessary. Reliability is about stability in measurements, where a procedure for measuring is reliable if it produces

similar results when applied to the same object. As an example, reliability matters a lot for questionnaires. Hornbæk and Law [70] showed how so-called homegrown questionnaires had lower reliability compared to carefully developed and validated questionnaires. Again, the recommendation is to use questionnaires developed in earlier work whenever possible. Reliability is also a concern when coding observations or categorizing outcomes of the process of interaction. Reliability cannot be assumed. Experiments that use dependent variables based on for instance observation should carefully define criteria for coding and have independent raters code the data and compare their coding (and report a measure of interrater reliability, see [38]). Oulasvirta et al. [112] contains an example of careful coding and reliability checking; their coding manual is also publicly available.

4.9 Describing the Interaction Process

The measures described in the previous section concern mostly the outcome of interaction; measures or descriptions of the *process of interaction* are also informative in much experimental work. One reason is that such descriptions help interpret and give context to variation in the dependent measures. A second reason is that they help speculate about potential mechanisms involved in producing changes (or lack thereof) in the dependent variables, the "why" of experiments. A third reason is that the HCI field knows too little about interaction processes: describing them may help advance our understanding of interaction (see Yi et al. [144] for an argument).

A widespread method for obtaining data that describe interaction is *logging*. Logging is an umbrella term for instrumenting user interfaces or environments to capture and store users' interaction, including time stamps and a means of relating interaction to the experiment's other variables. Typical data from logs include mouse movements/clicks, command usage, virtual navigation, and errors. Homegrown interfaces may write interaction events such as mouse movements and interface actions to a file or database; existing interfaces may be instrumented using software that captures keystrokes or mouse activity (e.g., Noldus's Observer, Techsmith's Morae); pre-instrumented user interfaces may be

used (e.g., web browsers that capture navigation activities such as [66]); or logs may be established from other sources (e.g., using Web logs, proxies). Lazar et al. [90] provided many additional examples of recent tools for logging. Further details on how to do and analyze logging are available in the literature (e.g., [65]). Independently of the specific tools used, the main point is that logging is low-cost and allows for detailed insights into interaction. Logging is just one of many ways to gather data about interaction. Other ways include videotaping interaction, capturing interaction by a screen recorder, tracking the movement of people, eye tracking, or finding traces of activity. Detailing these is out of the scope of the present work (see [90]).

Data on interaction may be summarized and analyzed using frequencies, sums, aggregates, co-occurrences of commands, transitions from one activity to another, detecting and comparing sequences of events, and many other ways. In addition to standard techniques for doing such analysis (e.g., [122]), approaches specific to HCI have been developed (e.g., [123]). Automated analyses during data collection may speed up things, but have to be planned in advance; subsequent exploration of patterns of interest may easily be done if logged data are placed in a database or otherwise readily available. To give one example of analysis of logging, Hornbæk and Frøkjær [69] used logs of mouse movements to create maps of how participants read scientific papers on a computer and to identify distinct phases of reading (i.e., for general understanding and for finding a specific piece of information).

A few comments about descriptions of the process of interaction are pertinent. First, it is worth iterating that while descriptions of interaction processes may work as dependent variables, they are not necessarily about quality in use. Hornbæk [67] gave several examples where this distinction was mixed up and authors concluded that one interface was better than another based on descriptions of the interaction process. Such a conclusion typically needs a warrant to explain why the description of the interaction process implies goodness or lack thereof. Second, describing the process of interaction is often key to unpacking and interpreting individual differences. They, in turn, are large and often important to understanding performance data. Third, capturing data on interaction is often unobtrusive, providing a number

of benefits as experimental evidence [138]. In a study of an experimental text viewer called TeSS [64], participants first used a variety of viewer features and then — rather than being asked about their preference — were simply given the choice of which viewer features to use for a final task. This choice were logged and provided an interesting, behavior-based indicator of preference. Fourth, the recommendation here is to log as much as possible and delay the decision on what to analyze. That tactic — in combination with current database capabilities — allows for more exploration and idea generation from interaction data. Remember, however, the earlier discussion of how to reason about and report ad hoc findings (compared to planned comparisons). An example of extensive logging is Henze et al. [63], who logged about 48 million keystrokes in a mobile phone game, including the exact place users tapped to produce the keystroke. They related these data to data collected in an ad hoc manner from the application store used to distribute the game (e.g., device used, screen size) and reported some interesting variations in tap distributions over devices.

Whereas experiments in HCI typically focus on quantitative data, many exemplary experiments also collect qualitative data, for instance in the form of interviews and observations. Some experiments also rely solely on qualitative data. For instance O'Hara and Sellen [109] reported a much-cited experiment on reading from paper and from a computer. While they used an experimental setup — using for instance random assignment of participants to either paper or a computer condition — they only reported qualitative data on reading strategies and activities that differed between paper and computer. Such data is valuable when experiments go well (as in O'Hara and Sellen's study), but it is also useful in understanding why an experiment failed.

5

Running Experiments

The design of an experiment prescribes most of its running. Having a *formal procedure* for the experiment is beneficial (e.g., [14]). A formal procedure may be a script that the experimenter follows, a piece of software that runs participants through the experiment, or another way of systematically instructing participants, administering treatments, collecting responses, and so forth. One benefit of a formal procedure is to ensure that participants experience the same instructions and advice from the experimenter, and thereby reduce variability. Another benefit stems from the observation that the procedure forms part of operationalizing the independent variables and of creating the experimental situation that we investigate [42]. Preparing the procedure in detail (and in writing) allows us to check and discuss whether or not it reflects our intention with the independent variable, for instance through reviewing it with peers.

Doing one or more *pilot studies* is important. A good pilot study tests the whole, or particular critical parts, of an experiment. The point is to test both the procedure, data collection (for instance, video recorders, logging software, observation templates), user interfaces, and

data analysis. Piloting data analysis can sometimes help identify serious omissions or mistakes in data collection.

Being present during an experiment *observing participants is highly useful*, if it is practically feasible and not expected to affect the experiment. One use is to capture data in a structured way, so as to characterize participants' behavior and supplement dependent variables. Such capturing may be done using structured coding schemes or open-ended observation notes. The design, collection, and analysis of these sources of observational data share the potential of other observation-based data collection, and may be thought about and designed as such (e.g., [8, 119]). One may also observe, not for systematic data collection and reporting, but for *building up an intuition about data* and the phenomena being studied. The purpose of such observation is to generate ideas about what to look for in data, to derive potential explanations of observations, and to identify surprising behavior. This latter use of observation has been highly useful, though time-consuming, in many experimental studies that I have been involved in. Be aware, of course, that as a designer of the experiment (and possibly creator of the conditions being compared) you have a vested interest in the outcome.

Doing a post-experimental follow up or *debriefing* is both useful to the experimenter and fair to participants. It gives a chance to take questions, hear comments, and explain the experiment to participants, if they desire to know more or need to understand why the experiment was designed in a particular way. Some participants also find it important to receive a copy of a final, written up report on the experiment; some may even have interesting comments on the experimenter's interpretation of their behavior. Note that doing experiments online or through app stores makes debriefing much harder; experimenters should be much more careful in such circumstances.

Finally, treating participants with *respect* is important. Respect implies recognizing participants as human beings, paying due regard to their needs, feelings, and well-being. It also implies not wasting participants' time on irrelevant or unimportant research. To me, being respectful to participants summarizes well the many concerns in doing ethical research. The basic principles of ethical research with humans were described in the Nuremberg Code and later in the Declaration

of Helsinki; later guidelines on ethics relevant to HCI include ACM Code of Ethics and Professional Conduct[1] and American Psychological Association's (APA) Ethical Principles of Psychologists and Code of Conduct.[2] Many HCI textbooks offer good advice on how to do ethical research (e.g., [90]). Some key things to ensure are as follows.

- Voluntary participation and informed consent. Participation in HCI experiments should be voluntary. It should be possible to opt out of the experiment at any time, without any repercussions. Experimenters should tell prospective participants about the purpose of the experiment and what participants will need to do, so that they can take an informed decision about whether to participate or not. In rare cases, a cover story or deception about the true nature of the experiment may be considered; in such cases particular care must be taken to ensure that the deception is necessary and that participants are debriefed.
- Protection from harm. Participants in HCI experiments should be protected from mental or physical harm. Consider both harm during and after the experiment.
- Privacy. Data from participants should be kept confidential and anonymous. Data from experiments should be stored so that outsiders cannot identify participants. Publications about the experiment should leave participants anonymous as far as possible; using photos and videos where participants are identifiable requires permission. Sharing and reporting data so that they cannot be related back to individuals are challenging, in particular if biometric, kinematic, georeferenced, or similar data are involved. Such data may be combined to identify participants or groups of participants, so be careful.
- Legal agreements and terms/conditions surrounding the experiment. Data collected through tools (e.g., social networks) must respect the terms and conditions of those tools.

[1] http://www.acm.org/about/code-of-ethics.
[2] http://www.apa.org/ethics.

The Association of Internet Researchers Ethics Guide[3] contains many ethical questions for researchers that collect data through web-based forums or social networks. Many countries have local data protection acts that should be followed. Also, some institutions have internal rules and processes that experimenters must follow (e.g., enforced by review boards, see below).

- Contact to researchers. Do a debriefing, as described above. It should be possible for participants to contact experimenters after the experiment with questions about the research. It should also be possible to opt out after the experiment, if a participant for some reason wants to. So give your contact details, also in online or app store studies.

Many institutions will have formal requirements in place to help experimenters do ethically defensible research. These requirements vary across institutions and countries. In Denmark, for instance, most experiments in HCI do not have to be formally registered and reviewed, except if they involve human or biological samples or if they relate to health care. In the US, many universities have set up institutional review boards (IRBs) that have to approve most or all experiments with humans. The IRB assesses ethics based on an application that may describe participants, procedure, risks and benefits of the research, steps to ensure confidentiality, materials for obtaining consent, and so forth. Lazar et al. [90] gave an example of an IRB application.

[3] http://ethics.aoir.org/.

6

Reporting Experiments

The way experiments are reported is crucial. Independently of how well you design and run experiments, insufficient care in reporting, statistics, or explanation of results may turn readers away or upset reviewers. Therefore, reporting experiments shares the difficulty of writing in general (see [11, 134] and many others). As a minimum, a research report must offer enough detail that the reader understand the design of the experiment and its results. Being clear is key. An example where experimental reports often fail is in poor, verbal descriptions of constructs [127]. In addition, a report needs to describe a design in sufficient detail that a reader can understand who the participants were, what constructs were hypothesized about, the experimental design, the key variables, how analysis were done, and how data were interpreted. Congruent with advice on writing in general, the reports should be succinct. There is no need to explain what statistics is about when one can refer to a textbook, nor is there any need to describe all data analyses considered. Assume the reader impatient with superfluous detail and long-windedness.

Supplementing this general advice, we next present four complementary heuristics for reporting experiments: justifying the design,

providing evidence, narrating results, and being open about alternative interpretations and limitations.

6.1 Justify the Design

As discussed in earlier sections, many difficult decisions must be made when designing and running experiments. These decisions often result from trade-offs, arguments adapted from earlier work, and reasoning about research methods. In addition to explaining the design, excellent reports on experiments therefore *justify the experimental design* to readers so that they understand why key decisions were taken. An excellent report explains most (or the most important) choices in experimental design, as they have been discussed in earlier parts of the present monograph. The following list may serve as an illustration:

- Explain hypotheses and how they are justified from earlier work. The paper mentioned earlier on why people find it annoying to overhear mobile phone conversations does this well [99]. It cites the earlier work and explains how it supports each of the hypotheses.
- Explain what the dependent variables are and why they were chosen. Pirhonen et al. [116] justified using a workload measure as follows: "Workload is important in a mobile setting as users must monitor their surroundings and navigate, therefore fewer attentional resources can be devoted to the computer. An interface that reduces workload is likely to be successful in a real mobile setting".
- Explain the tasks and their rationale. In a comparison of paper-based and computer-based documents, O'Hara and Sellen [109] wrote "we use an experimental task which we believe, based on our field studies of readings in organizations [reference to earlier work], is both naturalistic and representative of reading in real work settings".
- Choice of analysis: why were data analyzed this particular way and, if an unusual approach was taken, why not use a standard approach? When evaluating Bubble Cursor — a mouse interaction technique with a dynamic activation

> area — Grossman and Balakrishnan [53] wrote "A related issue was our decision not to use the effective width correction for accuracy" and explained why they used an alternative method for analysis, rather than the correction often used in the literature.

Justifying the design seems particularly relevant when general considerations on experimental design (e.g., on the benefits of a between-subjects design) interacts with, or even contradicts, insights from previous work or the subject matter of the experiment. The example from the paper on Bubble Cursor above illustrates this well [53]. Because a non-standard approach to analysis was decided upon, the authors took particular care in justifying that approach. Had they not, readers might have wondered about this decision.

The need to justify decisions in experimental design and analysis was expressed eloquently by Abelson [1, p. xii]:

> When you do research, critics may quarrel with the interpretation of your results, and you better be prepared with convincing counterarguments. (These critics may never in reality materialize, but the anticipation of criticism is fundamental to good research and data analysis. In fact, imagined encounters with antagonistic sharpsters should inform the design of your research in the first place.)

Although Abelson's sharpsters were brought up in the context of data analysis, they exist in similar numbers for research design. Explaining all key decisions in experimental design, preferable with explicit rationales and justification from earlier work, is for me an important check of the validity of a design. In addition to being invaluable in reporting experiments, such justifications are also a resource in designing an experiment. They help experimenters think through experiments.

6.2 Provide Evidence

Strong reports on experiments should provide evidence. For instance, the journal Nature instructs reviewers to check if a paper "provides

strong evidence for its conclusions"[1]; a recent instruction for submissions to ACM's CHI conference suggests that "the validity of your submission's contribution must be adequately supported by appropriate arguments, analyses, evaluations, or data as best fit the contribution type"[2]. How can you best provide evidence?

One idea is to *back up major conclusions with evidence*. Gray and Salzman [49] provided an illustrative study of how a selection of influential papers on usability evaluation failed to do so; they named this a lack of "conclusion validity". For instance, one paper would advise on which evaluation method to choose when resources are scarce, even though it had not studied such a situation. Conclusion validity may be ensured through careful writing. When we speculate about implications for design or advice to practitioners from an experiment, one should note so explicitly (rather than make readers believe that the speculations were shown by the experiment). Another way to ensure conclusion validity is to develop a clear chain of evidence. The notion of chain of evidence was discussed by Yin [145, p. 105] in the context of case studies (see [77] for a related notion). It suggests that researchers develop a clear and auditable account of how conclusions are derived from data. This notion is particularly important in case-study research, which integrates different types of data, often collected across different settings. But it also applies to experiments. A clear chain of evidence helps establish which conclusions may be drawn from a study and may be used to illustrate to a reader how key claims are backed up by data. One may, for instance, describe whether different ways of backing up a conclusion are consistent (e.g., across usability measures, other dependent variables, and data on interaction), explain why some participants performed unusually well or why performance on tasks differed (but were expected to show similar patterns), and check the representativeness of illustrations and of the key conclusions reported.

John and Marks [78] used a traceable chain of evidence well. They studied the usefulness of usability evaluation methods, in particular how trying to fix usability problems found with those methods affected

[1] http://www.nature.com/authors/editorial_policies/peer_review.html.
[2] http://chi2013.acm.org/authors/guides/guide-to-a-successful-archive-submission/.

the use of a revised interface: Would problems go away or be made worse by the fixes? Their study was documented with a so-called effectiveness tree, which illustrated all relations among usability problems, whether or not they had been corrected in the revised interface, and their effect on users. Even if the tree had not made it into the paper (and helped the reader), I imagine it was useful for John and Marks in organizing and checking their analysis.

A common way to provide evidence is with *descriptive and inferential statistics*. One purpose of such statistics is to summarize data; another purpose is to establish statistical conclusion validity, that is, to show that differences in numbers are not just random variation, but related to real differences between conditions. Many examples have been given of faults in statistical reasoning within HCI (e.g., [17, 79, 49]) and outside (e.g., [7]). Cairns [17], for instance, reviewed inferential statistics in a sample of papers on HCI and found that many reported insufficient details about the tests being performed, failed to check the assumptions of the tests, performed too many tests, or used the wrong tests. Many books, however, describe how to do and report statistics correctly, see for instance [57, 122, 141]; Abelson [1] covered principles of statistical thinking at a general level. Nevertheless a few comments specific to HCI may be given here.

Descriptive statistics is the easiest to deal with: its key purpose is to describe and summarize data. In my view, most papers fail here by inadequately describing data using descriptive statistics. When reading an experimental report, I expect that all key comparisons of an experiment are described in a clear way. Without descriptive data, statistical testing is uninformative. For every measure of central tendency (say, mean or median) one should also show variability (say, confidence interval or standard deviation). These pieces of information must be present independently of whether evidence is given in text, tables, or figures. Of course evidence may be communicated in many other ways than through text and numbers (e.g., [136]). Although not an experiment in the sense used in the present paper, Adar et al. [3] employed graphics brilliantly to illustrate differences in how users revisit web pages.

Inferential statistics concerns drawing conclusions from data with random variation, for instance due to sampling. Because almost all

experiments in HCI use samples of participants, they need to ensure that conclusions about differences between levels of the independent variables are valid and not just due to noise/randomness. Because of the sampling, one can almost never compare the means of task completion times for two user interfaces and conclude anything without some kind of inferential statistics. In HCI, inferential statistics is most often based on null-hypothesis significance testing (NHST). The NHST approach states a null hypothesis and uses particular tests — such as t-tests, χ^2 tests, analysis of variance (ANOVA), or Friedman's test — as evidence for an alternative hypothesis [107]. Although NHST is ubiquitous in HCI, many alternatives exist. For instance, it has been proposed to report only effect sizes in place of significance tests (e.g., [107]), to report only confidence intervals (e.g., [29]), or to use different modes of inference (such as Bayesian, see [107]). These proposals may readily be taken up by experimenters in HCI, although that is rarely done. Below we discuss mainly reporting based on NHST and adaptations of that approach (in particular as discussed in [141]). With that narrowing of statistical inference, two things make it easy to do statistics correctly. First, as mentioned above, how to choose and conduct statistical analyses is well covered in textbooks. Second, a set of simple analysis techniques will be sufficient for most experiments. Table 6.1 presents six commonly used tests in HCI and some things that experimenters should consider.

Statistical analysis is often quite complex, but there is much value in trying to simplify it as much as possible. One way to do so is to use simpler types of analysis (e.g., linear contrasts [121] instead of omnibus tests followed by post hoc tests), another is to use simpler experimental designs. It is worth keeping in mind this recommendation from a group of excellent statisticians [141]:

> Choosing a minimally sufficient analysis. The enormous amount of variety of modern quantitative methods leaves researchers with the nontrivial task of matching analysis and design to the research question. Although complex designs and state-of-the-art methods are sometimes necessary to address research questions effectively,

Table 6.1. Typical statistical tests used in HCI and some things to consider for each test.

Test	Purpose	Example	Things to consider
t-Test	Compares the means of two samples.	Alonso et al. [5] compared a table and a graphical layout. For individual tasks, they compared mean task completion time across layouts using unpaired t-tests, with $N-2$ degrees-of-freedom (where N is the number of participants in the experiment) and using a Bonferonni correction of α set to 0.05/31 (because 31 tasks were compared).	Choose between paired (samples from same persons, group, or experimental unit) and unpaired tests.
Analysis of Variance (ANOVA)	Compares the means of two or more samples.	Douglas et al. [31] used ANOVA to compare the pointing performance of participants using a joystick and a touchpad. One conclusion was that "The mean movement time for the joystick was 1.975 seconds with a standard deviation of 0.601 seconds. For the touchpad, the mean movement time was 2.382 seconds and a standard deviation of 0.802 seconds. These differences were statistically significant ($F_{1,22} = 11.223, p = 0.0029$)."	Many studies take the mean of performance for each participant in a condition and use them in the ANOVA (e.g., [31]). Check homogeneity of variance, a crucial assumption in ANOVA.
Repeated Measures ANOVA	Compares means over repeated measures from the same person.	Grossman and Balakrishnan [53] compared three cursor types (CT) and reported that "Repeated measures analysis of variance showed a significant main effect for CT ($F_{2,22} = 7947, p < 0.0001$)". They used repeated measures because CT was varied within-subjects and because four other independent variables were manipulated.	The repeated measures may be means of a person's performance in a condition (as for ANOVA).

(*Continued*)

Table 6.1. (*Continued*)

Test	Purpose	Example	Things to consider
Multivariate Analysis of Variance (MANOVA)	Compares samples with multiple dependent variables.	Frandsen-Thorlacius et al. [40] compared differences in the perception of usability among cultures on seven scales of usability. They did one overall MANOVA of the hypothesis that perception differs between cultures and followed up with ANOVAs of each of the scales.	MANOVA has many assumptions that are crucial to check (e.g., multivariate normality).
Non-parametric t-tests and ANOVAs	Tests that work as above but does not assume a normal distribution of data.	Forlines et al. [39] compared the use of mouse and touch input on a tabletop display using a within-subjects design. They used a Wilcoxon signed-rank test to compare preference data.	Useful for analyzing ratings (e.g., preference). Non-parametric ANOVA tailored for HCI are available [80, 143]. Non-parametric tests have lower power and make it hard to deal with interactions.
χ^2	Tests the difference in frequencies across categories.	Guan et al. [55] compared retrospective think aloud testing across two levels of task complexity. They categorized participants' verbal behavior and compared the frequency across across task complexity with a χ^2 test.	Assumes no ordering among categories. Must have a sufficient number of observations/expected observations in each cell.

simpler classical approaches often can provide elegant and sufficient answers to important questions.

In addition to the basic reasoning above and in Table 6.1, the following list presents issues about inferential statistics that are frequently misunderstood or overlooked in HCI (see also [33, 79]).

- In some papers (and in many textbooks) a null hypothesis is discussed: a null hypothesis is simply assuming no difference between conditions (although the real/substantive hypothesis is that there is a difference). The null hypothesis is a statistical trick that need not be mentioned in reports on experiments and that cannot ever be said to be accepted (because it was assumed to begin with).
- Low p-values from a statistical test do not mean importance, nor do they mean generalizable or indicate the likelihood of the (real) hypothesis being investigated. A p-value merely represents the probability — given or assuming the null hypothesis — that the samples come from the same distribution (i.e., that they are similar). The key here is to understand that "given or assuming" is crucial. The probability of data given the null hypothesis (what we do in significance testing) is different from the probability of the null hypothesis given the data (what one may mistakenly do in taken p-values to indicate the likelihood of the hypothesis). Confusing these probabilities is called the fallacy of the transposed conditional and may be illustrated by the difference in the probability of being dead given that one has been lynched (high) and the probability of having been lynched given one is dead (low) [79]. Nickerson [107] discussed other potential misunderstanding of p-values.
- Statistical significance is different from magnitude (and the latter is often more important). Effect size quantifies magnitude and is easy to compute (e.g., [122]); magnitude may also be illustrated by comparison of effects to earlier studies or known standards of performance [1]. Effect size also has the

nice property that it is unaffected by sample size, in contrast to p-values.

- If one engages in significance testing, one should not focus (too much) on marginal significant results or on results where the significance level was fixed only at the time of analysis. Doing so undermines the logic of statistical significance testing.
- The notion of experiment-wide (or family-wise) error refers to the risk, across a set of hypotheses investigated in an experiment, to accept one as significant when it is not. It is often invoked by reviewers when authors conduct a lot of statistical tests and emphasize a few, significant ones. The key issue is that if one does multiple tests, the chance of accepting one as significant by mistake increases. For a single test, the chance is given by the cutoff for significance, often referred to as α, typically 0.05. For n tests it is given by $1 - (1 - \alpha)^n$: if you do 10 tests, the real chance of accepting a result as significant by mistake is 40%. One often used remedy is to use an omnibus analysis of variance; a second remedy is to use targeted tests such as contrast analysis [121]; a third remedy is to use post-hoc tests, for instance using the Bonferroni correction (which uses α/n as criterion for rejection).
- It is notoriously difficult to use an experiment failing to find a difference to conclude anything. At the very least, the power of the experiment should be calculated and used to justify why readers may learn something from a lack of result (recall that power refers to the likelihood of detecting a difference that is present). Sonnenwald et al. [132] used power analysis to make clear to readers what could reasonably be concluded from the null result of a comparison of scientific collaboratories to a control condition.
- Leaving out participants or other subsets of data on an ad-hoc basis violates the logic of statistical testing. It is also a no-go to add participants to an experimental setup if one has already begun analyzing it: Simmons and colleagues [130] showed how this practice inflates the rate of false-positive results. They showed how researchers can find significant,

but false, effects 22% of the time (instead of 5%) by starting with 10 participants per condition, and continue to add participants and test after every participant added. A strict criterion should be used to eliminate individual data points or all data from a participant as outliers. Many such criteria exist, including trimming distributions in either end (say, by 2.5%), removing values more than three standard deviations from the mean, or discarding observations that are more than three interquartile ranges from the median or more than 1.5 interquartile ranges from the 25% and 75% quartiles (where quartiles is the set of three points that divide the data into four equally sized groups).

- Any experimental report that uses inferential statistics should check the degrees of freedom in its tests. Although degrees of freedom is explained in most textbooks, some experimenters get it wrong, raising doubts about whether the right tests have been performed. For instance, F-tests are typically reported with degrees of freedom for the numerator and denominator of the statistic (also called the between-group and within-group/error degree of freedom, respectively). If one do a between subjects experiment with 3 interfaces and 10 persons using each interface (total of 30), the degrees of freedom of an ANOVA should be $3 - 1 = 2$ (for the numerator or between-groups term) and $30 - 3 = 27$ (denominator, within-group). If one instead do a within-subjects experiment, also with 3 interfaces and 10 persons in total, then the degrees of freedom of the ANOVA should be $3 - 1 = 2$ and $(3 - 1) * (10 - 1) = 18$. And if one opts to do more complex statistics, different denominators might be used (with different degrees of freedom).
- All statistical tests involve assumptions. Know them and check them. One assumption sometimes ignored in HCI papers is that observations should be able to vary independently (i.e., be statistically independent). For instance, if participants worked in a group, the inferential statistics should be done on the group level because measures of individual

participants' performance are not independent or one should deal with the group effect in other ways (i.e., with more complex statistics).
- Tests may be one- or two-sided; I remain unconvinced of most one-sided tests in HCI and recommend using two-sided tests. In a survey of the use of one-sided tests in two journals on ecology and animal behavior, Lombardi and Hulbert [92] concluded that "all uses of one-tailed tests in the journals surveyed seem invalid" (p. 447).

Although this list seems daunting, errors in statistics can normally be fixed by choosing the right analysis and reasoning correctly about what an analysis shows; rarely do such errors make it necessary to rerun an experiment.

6.3 Narrate Results for the Reader

A key quality of excellent reports on experiments is *narration*. In the words of Abelson [1], "Meaningful research tells a story with some point to it" (p. xiii). One way of thinking about narration is to think about the essential point an experiment is making. Key to this is becoming clear about the contribution, highlighting it in an experimental report, and discussing a potential reader's questions about it. The goal of narration also implies that readers are interested in understanding the results and potential explanations thereof, and rarely in statistics or numbers in themselves. Also, narration is not about retelling the order in which the experiment was designed or the manner in which results were obtained: it is about telling a coherent and clear story.

Another purpose of narration is to make the reader understand *why results are interesting*. This is about answering a potential sharpster's questions that an experiment contains "no surprises". Some prominent ways of arguing why a result is interesting are as follows (see Davis [30] for other ways to argue interestingness):

- To compensate for deficiencies in earlier work.
- To show something for the first time.
- To disprove something believed to be true.

- To reconcile two conflicting views.
- To develop better theories or models of interactive behavior.

One example of such an argument is seen in a paper by Zhai and colleagues [147]. They argued that an earlier study of target expansion was inadequate and designed an experiment to set those limitations right.

The position taken here is that this is the experimenter's responsibility to argue interestingness. Any contribution must be described in the context of earlier work if readers are to figure out why it is interesting. Wilkinson and the task force on statistical inference [141] put this nicely:

> Do not interpret a single study's results as having importance independent of the effects reported elsewhere in the relevant literature. The thinking presented in a single study may turn the movement of the literature, but the results in a single study are important primarily as one contribution to a mosaic of study effects.

In narrating results, most readers want to know "why", that is, thoughts on the part of the experimenter on *mechanisms behind the observed results*. When explaining mechanisms, interaction patterns and qualitative data about behavior may supplement the data from dependent variables. Models may serve as one such explanation; simple, explanatory concepts or examples of prototypical behavior may provide others.

As a tactic for narration, Abelson [1] introduced the notion of ticks, buts, and blobs. A *tick* is a detailed statement of a distinct research result (as in ticking off on a form); a *but* is a statement that qualifies or constrains ticks (which we will discuss in more detail in the next section); and a *blob* is a "cluster of undifferentiated research results" ([1, p. 105]). For Abelson, a tick represents an important finding that adds to the field. It is not to be confused with omnibus test results (e.g., an F-test showing a significant difference among three conditions), a non-significant test (unless strong power and prior work suggest

otherwise), or an uninteresting finding (e.g., that width influence performance in Fitts' law experiments). A tick is also different from the blob paragraph, which lists many significant tests but never makes it clear to the reader which are important additions to the research field and which are just accidental or unimportant. Good experimental reports provide few, but clearly articulated ticks.

6.4 Acknowledge Alternative Interpretations and Limitations of Results

As mentioned in an earlier section, experiments have the potential to limit the influence of "the ego-centric fallacy", as well as other biases in the experimenter's judgment (such as confirmation bias, see [93]). Most scientists also realize that any single experiment is insufficient. Reports on experiments should therefore (a) discuss alternative interpretations of data and their relative merits and (b) openly acknowledge limitations and concerns in the collection or interpretation of results. If possible, also present potential remedies to limitations and concerns.

Data from experiments and the analysis of them into some tentative conclusion are never straightforward, although published papers sometimes present it that way. The advice here is to carefully *acknowledge and discuss alternative interpretations of data*. This may both address readers' questions about data and interpretation hereof, and may help focus subsequent research. A couple of approaches to doing so may be given. First, in an earlier section we discussed articulation and the "but". "Buts" are an excellent approach to acknowledging alternative interpretations. Second, working through and discussing alternative interpretations works well. As an example, Hornbæk and Law [70] studied correlations among usability measures. They found that correlations were medium to small. The discussion of that finding was structured around two alternative interpretations: that it confirmed earlier work and that it did not match the expectations raised in earlier work. Hornbæk and Law discussed these two interpretations and the associated earlier work before attempting to conclude on the data.

All *experimental results are limited in one way or the other*. Wilkinson et al. [141] gave a nice summary of how and why to

acknowledge limitations. They wrote: "Note the shortcomings of your study. Remember, however, that acknowledging limitations is for the purpose of qualifying results and avoiding pitfalls in future research. Confession should not have the goal of disarming criticism." Limitations may concern many parts of the experiment, including its design, what actually happened when running it, and results that do not make sense given earlier work. Discussing these is key to great reports on experiments. Accot and Zhai [2], for instance, followed their set of experiments that helped derive the steering law with the remark that "It should be pointed out, however, that there are various limitations to these simple laws" (p. 301) and discussed how more work is needed to understand the impact of body limitations and handedness, and to achieve a higher level of generality.

One reason for highlighting the need to acknowledge limitations is that the pressure to publish reports on experiments may lead experimenters to leave out important information. Simmons et al. [130] showed how experimenters could obtain nicely looking results by leaving out data, dropping dependent measures, and attempting different analysis approaches. It is clear that such approaches are detrimental to research. Experimenters should report limitations openly.

7

Pragmatics of Experiments

The previous sections have discussed some ideals for the design, running, and reporting of experiments. Next, I briefly want to discuss some issues in doing real experiments.

Some experiments, perhaps the majority, fail. They do not produce the hypothesized differences, a task turns out to be irrelevant, or individual differences are too large to allow any conclusions. At the design stage of an experiment one may therefore sensibly *think of a fallback plan*: A plan to salvage some of the resources put into an experiment if it should fail. Such a plan could entail collecting extra data that might be of interest independently of the outcome of the experiment (qualitative data or log data often work well in this regard) or adding an additional level of independent variable (as discussed earlier), allowing for discussion of variations in the independent variable. Piloting experiments may also reduce the risk of failure (or allow an experimenter to abandon or rethink a particular experimental design).

Many parts of this monograph has mentioned simplification, in particular with respect to the design of experiments and with respect to the analysis of results. In being pragmatic about experimentation in HCI, an important insight is that *one experiment cannot do all*. Most

often it makes sense to tease apart intricate designs and complex procedures to create simple and understandable experiments. One may think of a series of smaller experiments, rather than one, all-encompassing experiment. For instance, Cockburn and colleagues [24] studied menu navigation and ran a small calibration study and a study of real menu designs, rather than just doing one large study. Thereby the empirical part of their argument became more step-wise and provided more room for correcting poor choices in experimental design and theoretical modeling.

As discussed in earlier parts of this monograph, experiments are good for certain research questions and poor for others. One practical challenge in doing experiments is to ensure that the *experimental work is aligned with the research question.* The work should differ depending on whether one wants to study if UI1 works better than UI2, if a phenomenon is possible with a UI, if a UI works in real life, if feature X works better than feature Y, or some other question. We have already mentioned many potential trade-offs that experimenters face when trying to align experimental work and research question: control vs realism, simplification vs complexity, existence proof vs mechanisms, and systems vs techniques; others include utility of systems vs usability, ambition vs costs, and publishability vs potential impact. Let us revisit one such tradeoff, the extent to which variability should be controlled, to illustrate the trade-offs to be made. With variability we mean variation in dependent measures that are unrelated to the independent variables. In most cases, we aim to reduce such variability because it decreases statistical power [127]. Such control could imply using blocks in an experiment, increasing the reliability of measures, reducing random variability from the setting, ensuring a consistent understanding of instructions, and so forth. Nevertheless, as discussed in the initial sections of this work, control typically hurts realism: controlling tasks means not getting the insight due to variation in actual tasks users would want to do; controlling the location of use means missing impromptu adaptation of user interfaces to a location-dependent need or opportunity. This may or may not be a problem, depending on the research goals of the experiment. The trade-off here is that control of variability comes at the expense of lowering realism and

potentially missing interesting interaction behavior: thinking through this issue and finding an appropriate ambition for an experiment is a tough practical question for experimenters. Many other such trade-offs should be thought about for the particular research question that one is addressing. Again, one experiment cannot do it all.

One important contribution experimenters can make is to share the materials of their experiment publicly. Such materials could include user interfaces, description of procedures, tasks, the data collected, and the statistical analysis. Inexperienced experimenters can learn a lot from such material and it enables independent scrutiny (and even replications) of experiments. One excellent example of such a contribution was made by Jansen et al. [74]. They studied physical visualizations and in addition to the published paper, detailed information is available online about the visualizations, the tasks, the data, the logging, and the data analysis. Sharing such material with the scientific community is very valuable.

8

Conclusion

Experiments in HCI work by deliberately introducing interventions that might affect the interaction between humans and computers and describing the effects. They form an important part of HCI methodology. The present monograph has described some heuristics for designing, running, and reporting experiments. We have argued and sought to exemplify how the quality of experiments in HCI can been improved through the use of the heuristics. The heuristics are summarized in Table 3.1. In particular, we suggested to design experiments that are focused on a clear research question and pursue strong comparisons. Earlier work has been argued invaluable in designing and reporting experiments. Reports on experiments should offer evidence and take care in narrating results; they should also explore alternative interpretations of results and discuss limitations of the experimental work. Participants in experiments should be treated with respect and their understanding of the experimental situation should be given careful attention. We have also argued that any experiment is limited and that experimenters should consider this fact when designing and reporting experiments.

In concluding, I want to remove a potential misunderstanding about the goal of experiments. Many of the preceding discussions have been about presenting experiments; perhaps some of those discussions have conveyed the impression that good experiments are only about being able to describe and justify an experimental design or about being able to package results neatly. They are not. Contribution, soundness, substance, and perhaps even being right are also important characteristics. As Giner-Sorolla [46] has argued, the pressure to publish and the increased competitiveness in widely read research outlets make aesthetics in research matter more. Aesthetics in research is about making designs, data, and results clear and pleasing; it is the opposite of the messiness, complexity, and open questions that many careful experimenters experience. Aesthetics, in Giner-Sorolla's sense, leads to more emphasis on novelty and importance to ongoing discussions in a research field, and to less emphasis on scientific soundness. It seems a dilemma. On the one hand, experiments should get published and be presented in a clear and pleasing way. On the other hand, experiments are about doing the right thing, about seeking out complexity, and about taking risks. If anything, the present work has emphasized the latter. Nevertheless, every experimenter in HCI must balance the horns of this dilemma by carefully fleshing out and trading off the whys and hows of experimental work.

Acknowledgments

Thanks to Aran Lunzer, who inspired me to write this paper. I am grateful for comments on earlier drafts by Ben Bederson, Morten Hertzum, Harry Hochheiser, Alex Tuch, Shumin Zhai, and several anonymous reviewers.

References

[1] R. Abelson, *Statistics as Principled Argument*. New York: Lawrence Erlbaum, 1995.

[2] J. Accot and S. Zhai, "Beyond Fitts' law: models for trajectory-based HCI tasks," in *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pp. 295–302, New York, NY, USA: ACM, 1997.

[3] E. Adar, J. Teevan, and S. T. Dumais, "Large scale analysis of web revisitation patterns," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1197–1206, New York, NY, USA: ACM, 2008.

[4] R. Adcock and D. Collier, "Measurement Validity: A shared Standard for Qualitative and Measurement Validity: A shared Standard for Qualitative and Quantitative Research.," *American Political Science Review*, vol. 95, no. 3, pp. 529–546, 2001.

[5] D. L. Alonso, A. Rose, C. Plaisant, and K. L. Norman, "Viewing personal history records: A comparison of tabular format and graphical presentation using LifeLines," *Behaviour & Information Technology*, vol. 17, no. 5, pp. 249–262, 1998.

[6] C. A. Anderson, J. J. Lindsay, and B. J. Bushman, "Research in the Psychological Laboratory: Truth or Triviality?," *Current Directions in Psychological Science*, vol. 8, no. 1, pp. 3–9, 1999.

[7] J. Bailar and F. Mosteller, "Guidelines for statistical reporting in articles for medical journals," *Annals of Internal Medicine*, vol. 108, no. 2, pp. 266–73, 1998.

[8] R. Bakeman, "Behavioral observation and coding," in *Handbook of research methods in social and personality psychology*, (H. T. Reis and C. M. Judd, eds.), pp. 138–159, Cambridge, UK: Cambridge University Press, 2000.

[9] J. A. Bargas-Avila and K. Hornbæk, "Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2689–2698, New York, NY, USA: ACM, 2011.

[10] L. Barkhuus and J. A. Rode, "From Mice to Men - 24 Years of Evaluation in CHI," in *Paper presented at alt.chi*, 2007.

[11] J. Barzun, *Simple & Direct*. New York, NY: Harper Perennial, 2001.

[12] L. Berkowitz and E. Donnerstein, "External validity is more than skin deep: Some answers to criticisms of laboratory experiments.," *American psychologist*, vol. 37, no. 3, p. 245, 1982.

[13] N. Bevan, "Measuring usability as quality of use," *Software Quality Journal*, vol. 4, no. 2, pp. 115–130, 1995.

[14] A. Blandford, A. Cox, and P. Cairns, "Controlled experiments," in *Research methods for Human Computer Interaction*, (P. Cairns and A. L. Cox, eds.), Cambridge, UK: Cambridge University Press, 2008.

[15] M. Bunge, *Causality and modern science*. New York: Dover Publications, 3rd ed., 1979.

[16] M. M. Burnett, L. Beckwith, S. Wiedenbeck, S. D. Fleming, J. Cao, T. H. Park, V. Grigoreanu, and K. Rector, "Gender pluralism in problem-solving software," *Interacting with Computers*, vol. 23, no. 5, pp. 450 – 460, 2011.

[17] P. Cairns, "HCI... not as it should be: inferential statistics in HCI research," in *Proceedings of the 21st British HCI Group Annual Conference on People and Computers*, pp. 195–201, Swinton, UK: British Computer Society, 2007.

[18] J. Campbell, "Labs, fields, and straw issues," in *Generalizing from laboratory to field settings: Research findings from industrial-organizational psychology, organizational behavior, and human resource management*, (E. Locke, ed.), pp. 269–279, Lexington, MA: Lexington Books, 1986.

[19] J. Campbell, R. Daft, and C. Hulin, *What to study: Generating and developing research questions*. Beverly Hills, CA: Sage, 1982.

[20] S. K. Card, W. K. English, and B. Burr, "Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT," *Ergonomics*, vol. 21, no. 8, pp. 601–613, 1978.

[21] S. Carpendale, "Evaluating information visualizations," in *Information Visualization: Human-Centered Issues and Perspectives*, (A. Kerren, J. Stasko, J.-D. Fekete, and C. North, eds.), pp. 19–45, Berlin: Springer, 2008.

[22] T. Chamberlin, "The method of multiple working hypotheses," *Science*, vol. 15, no. 366, pp. 92–96, reprinted 1965, v. 148, p. 754-759., 1890.

[23] J. P. Chin, V. A. Diehl, and K. L. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 213–218, New York, NY, USA: ACM, 1988.

[24] A. Cockburn, C. Gutwin, and S. Greenberg, "A predictive model of menu performance," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 627–636, New York, NY, USA: ACM, 2007.

[25] A. Cockburn and B. McKenzie, "Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 203–210, New York, NY, USA: ACM, 2002.

[26] J. Cohen, *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum, 1988.

[27] J. Cohen, "A power primer," *Psychological bulletin*, vol. 112, no. 1, p. 155, 1992.

[28] H. Cooper, *Synthesizing research: A guide for literature reviews*. Thousand Oaks, CA: Sage Publications, third ed., 1998.

[29] G. Cumming and S. Finch, "Inference by eye: Confidence intervals and how to read pictures of data," *American Psychologist*, vol. 60, pp. 170–180, 2005.

[30] M. Davis, "That's interesting," *Philosophy of the Social Sciences*, vol. 1, no. 2, p. 309, 1971.

[31] S. A. Douglas, A. E. Kirkpatrick, and I. S. MacKenzie, "Testing pointing device performance and user assessment with the ISO 9241, Part 9 standard," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 215–222, New York, NY, USA: ACM, 1999.

[32] P. Dragicevic, A. Bezerianos, W. Javed, N. Elmqvist, and J.-D. Fekete, "Temporal distortion for animated transitions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2009–2018, New York, NY, USA: ACM, 2011.

[33] M. Dunlop and M. Baillie, "Paper Rejected ($p > 0.05$): An Introduction to the Debate on Appropriateness of Null-Hypothesis Testing," *International Journal of Mobile Human Computer Interaction (IJMHCI)*, vol. 1, no. 3, pp. 86–93, 2009.

[34] P. Edwards, F. Sainfort, T. Kongnakorn, and J. Jacko, "Methods of Evaluating Outcomes," in *Handbook of Human Factors and Ergonomics*, (G. Salvendy, ed.), pp. 1150–1187, Hoboken, NJ: Wiley, third ed., 2006.

[35] D. E. Egan, J. R. Remde, L. M. Gomez, T. K. Landauer, J. Eberhardt, and C. C. Lochbaum, "Formative design evaluation of superbook," *ACM Transactions on Information Systems*, vol. 7, no. 1, pp. 30–57, Jan 1989.

[36] D. Fanelli, "Negative results are disappearing from most disciplines and countries," *Scientometrics*, vol. 90, no. 3, pp. 891–904, 2012.

[37] G. Firebaugh, *Seven rules for social research*. Princeton, NJ: Princeton University Press, 2008.

[38] J. Fleiss, *Statistical methods for rates and proportions*. New York, NY: John Wiley & Sons, second ed., 1981.

[39] C. Forlines, D. Wigdor, C. Shen, and R. Balakrishnan, "Direct-touch vs. mouse input for tabletop displays," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 647–656, New York, NY, USA: ACM, 2007.

[40] O. Frandsen-Thorlacius, K. Hornbæk, M. Hertzum, and T. Clemmensen, "Non-universal usability?: a survey of how usability is understood by Chinese and Danish users," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 41–50, New York, NY, USA: ACM, 2009.

[41] E. Frøkjær, M. Hertzum, and K. Hornbæk, "Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 345–352, New York, NY, USA: ACM, 2000.

[42] H. L. Fromkin and S. Streufert, "Laboratory Experimentation," in *Handbook of Industrial and Organizational Psychology*, (M. Dunnette, ed.), Chicago, IL: Rand-McNally, 1976.

[43] J. Garst, N. Kerr, S. Harris, and L. Sheppard, "Satisficing in hypothesis generation," *The American journal of psychology*, vol. 115, no. 4, pp. 475–500, 2002.

[44] V. Gawron, *Human performance, workload, and situational awareness measures handbook*. Boca Raton, FL: CRC, 2008.

[45] A. S. Gerber and D. P. Green, *Field experiments: Design, analysis, and interpretation*. New York, NY: WW Norton, 2012.

[46] R. Giner-Sorolla, "Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science," *Perspectives on Psychological Science*, vol. 7, no. 6, pp. 562–571, 2012.

[47] R. L. Glass, V. Ramesh, and I. Vessey, "An analysis of research in computing disciplines," *Communications of the ACM*, vol. 47, no. 6, pp. 89–94, Jun 2004.

[48] J. D. Gould, J. Conti, and T. Hovanyecz, "Composing letters with a simulated listening typewriter," *Communications of the ACM*, vol. 26, no. 4, pp. 295–308, Apr 1983.

[49] W. Gray and M. Salzman, "Damaged merchandise? A review of experiments that compare usability evaluation methods," *Human–Computer Interaction*, vol. 13, no. 3, pp. 203–261, 1998.

[50] S. Greenberg and B. Buxton, "Usability evaluation considered harmful (some of the time)," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 111–120, New York, NY, USA: ACM, 2008.

[51] A. Greenwald, "Within-subjects designs: To use or not to use?," *Psychological Bulletin*, vol. 83, no. 2, p. 314, 1976.

[52] A. Greenwald, A. Pratkanis, M. Leippe, and M. Baumgardner, "Under what conditions does theory obstruct research progress?," *Psychological review*, vol. 93, no. 2, p. 216, 1986.

[53] T. Grossman and R. Balakrishnan, "The bubble cursor: enhancing target acquisition by dynamic resizing of the cursor's activation area," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 281–290, New York, NY, USA: ACM, 2005.

[54] T. Grossman, G. Fitzmaurice, and R. Attar, "A survey of software learnability: metrics, methodologies and guidelines," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 649–658, New York, NY, USA: ACM, 2009.

[55] Z. Guan, S. Lee, E. Cuddihy, and J. Ramey, "The validity of the stimulated retrospective think-aloud method as measured by eye tracking," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1253–1262, New York, NY, USA: ACM, 2006.

[56] C. Gutwin and S. Greenberg, "Effects of awareness support on groupware usability," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 511–518, New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1998.

[57] P. Harris, *Designing and reporting experiments in psychology*. Milton Keynes: Open University Press, 2008.

[58] S. Hart and L. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Human mental workload*, (P. A. Hancock and N. Meshkati, eds.), pp. 139–183, North Holland: Elsevier, 1988.

[59] M. Hassenzahl, "Experience Design: Technology for all the right reasons," *Synthesis Lectures on Human-Centered Informatics*, vol. 3, no. 1, pp. 1–95, 2010.

[60] M. Hassenzahl, M. Burmester, and F. Koller, "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität," in *Mensch & computer*, pp. 187–196, 2003.

[61] M. Hassenzahl and A. Monk, "The inference of perceived usability from beauty," *Human–Computer Interaction*, vol. 25, no. 3, pp. 235–260, 2010.

[62] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world," *Behavioral and Brain Sciences*, vol. 33, no. 2-3, pp. 61–83, 2010.

[63] N. Henze, E. Rukzio, and S. Boll, "Observational and experimental investigation of typing behaviour using virtual keyboards for mobile devices," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pp. 2659–2668, New York, NY, USA: ACM, 2012.

[64] M. Hertzum and E. Frøkjær, "Browsing and querying in online documentation: a study of user interfaces and the interaction process," *ACM Transactions on Computer-Human Interaction*, vol. 3, no. 2, pp. 136–161, Jun 1996.

[65] D. M. Hilbert and D. F. Redmiles, "Extracting usability information from user interface events," *ACM Computing Surveys*, vol. 32, no. 4, pp. 384–421, Dec 2000.

[66] J. I. Hong and J. A. Landay, "WebQuilt: a framework for capturing and visualizing the web experience," in *Proceedings of the 10th international conference on World Wide Web*, pp. 717–724, New York, NY, USA: ACM, 2001.

[67] K. Hornbæk, "Current practice in measuring usability: Challenges to usability studies and research," *International Journal of Man-Machine Studies*, vol. 64, no. 2, pp. 79–102, 2006.

[68] K. Hornbæk, "Dogmas in the assessment of usability evaluation methods," *Behaviour & Information Technology*, vol. 29, no. 1, pp. 97–111, 2010.

[69] K. Hornbæk and E. Frøkjær, "Reading patterns and usability in visualizations of electronic documents," *ACM Transactions on Computer-Human Interaction*, vol. 10, no. 2, pp. 119–149, 2003.

[70] K. Hornbæk and E. L.-C. Law, "Meta-analysis of correlations among usability measures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 617–626, New York, NY, USA: ACM, 2007.

[71] W. Hwang and G. Salvendy, "Number of people required for usability evaluation: the 10±2 rule," *Communications of the ACM*, vol. 53, no. 5, pp. 130–133, 2010.

[72] ISO, *ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability*. Geneva, Switzerland: International Organization for Standardization, 1998.

[73] M. R. Jakobsen and K. Hornbæk, "Evaluating a fisheye view of source code," in *Proceedings of the 2006 Conference on Human Factors in Computing Systems*, pp. 377–386, New York, NY: ACM Press, 2006.

[74] Y. Jansen, P. Dragicevic, and J.-D. Fekete, "Evaluating the Efficiency of Physical Visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2593–2602, New York, NY, USA: ACM Press, 2013.

[75] S. Jarvenpaa, G. Dickson, and G. DeSanctis, "Methodological issues in experimental IS research: experiences and recommendations," *MIS quarterly*, vol. 9, no. 2, pp. 141–156, 1985.

[76] R. Jeffries, J. Miller, C. Wharton, and K. Uyeda, "User interface evaluation in the real world: a comparison of four techniques," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 119–124, New York, NY: ACM, ACM Press, 1991.

[77] B. John, "Evidence-based practice in human-computer interaction and evidence maps," *ACM SIGSOFT Software Engineering Notes*, vol. 30, no. 4, pp. 1–5, 2005.

[78] B. John and S. Marks, "Tracking the effectiveness of usability evaluation methods," *Behaviour & Information Technology*, vol. 16, no. 4-5, pp. 188–202, 1997.

[79] M. Kaptein and J. Robertson, "Rethinking statistical analysis methods for CHI," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pp. 1105–1114, New York, NY, USA: ACM, 2012.

[80] M. C. Kaptein, C. Nass, and P. Markopoulos, "Powerful and consistent analysis of likert-type ratingscales," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2391–2394, New York, NY, USA: ACM, 2010.

[81] D. Kelly, "Methods for evaluating interactive information retrieval systems with users," *Foundations and Trends in Information Retrieval*, vol. 3, no. 1-2, pp. 1–224, 2009.

[82] G. Keren, "Between-or within-subjects design: A methodological dilemma," in *A Handbook for Data Analysis in the Behaviorial Sciences*, (G. Keren and C. Lewis, eds.), Hillsdale, NJ: Lawrence Erlbaum, 1992.

[83] N. Kerr, "HARKing: Hypothesizing after the results are known," *Personality and Social Psychology Review*, vol. 2, no. 3, pp. 196–217, 1998.

[84] J. Kjeldskov, M. Skov, B. Als, and R. Høegh, "Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field," in *Mobile Human-Computer Interaction*, pp. 529–535, Springer, 2004.

[85] H. Lam and T. Munzner, "Increasing the utility of quantitative empirical studies for meta-analysis," in *Proceedings of the 2008 conference on BEyond time and errors: novel evaLuation methods for Information Visualization*, pp. 21–27, ACM, 2008.

[86] T. Landauer, "Research methods in human-computer interaction," in *Handbook of human-computer interaction*, (M. Helander, T. K. Landauer, and P. Prabhu, eds.), pp. 203–227, Amsterdam: Elsevier, second ed., 1997.

[87] P. Lang, "Behavioral treatment and bio-behavioral assessment: Computer applications," in *Technology in Mental Health Care Delivery Systems*, (J. Sidowski, H. Johnson, and T. Williams, eds.), Norwood, NJ: Ablex, 1980.

[88] K. Larson and M. Czerwinski, "Web page design: implications of memory, structure and scent for information retrieval," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 25–32, New York, NY, USA: ACM Press, 1998.

[89] T. Lavie and N. Tractinsky, "Assessing dimensions of perceived visual aesthetics of web sites," *International Journal of Human-Computer Studies*, vol. 60, no. 3, pp. 269–298, 2004.

[90] J. Lazar, J. Feng, and H. Hochheiser, *Research methods in human-computer interaction*. Chichester, UK: Wiley, 2010.

[91] J. Lewis, "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use," *International Journal of Human-Computer Interaction*, vol. 7, no. 1, pp. 57–78, 1995.

[92] C. M. Lombardi and S. H. Hurlbert, "Misprescription and misuse of one-tailed tests," *Austral Ecology*, vol. 34, no. 4, pp. 447–468, 2009.

[93] R. MacCoun, "Biases in the interpretation and use of research results," *Annual review of psychology*, vol. 49, no. 1, pp. 259–287, 1998.

[94] S. Mahlke and M. Thüring, "Studying antecedents of emotional experiences in interactive contexts," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 915–918, New York, NY: ACM, ACM Press, 2007.

[95] D. Martin, *Doing psychology experiments*. Wadsworth Publishing Company, 2007.

[96] J. E. McGrath, "Methodology matters: Doing research in the behavioral and social sciences," in *Human-Computer Interaction: Toward the Year 2000*, (R.M.Baecker, J. Grudin, and W. A. Buxton, eds.), pp. 152–169, Los Altos, CA: Morgan Kaufmann Publishers, 1995.

[97] W. McGuire, "Creative hypothesis generating in psychology: Some useful heuristics," *Annual Review of Psychology*, vol. 48, no. 1, pp. 1–30, 1997.

[98] D. Meister, *Conceptual aspects of human factors*. Johns Hopkins University Press Baltimore, 1989.

[99] A. Monk, J. Carroll, S. Parker, and M. Blythe, "Why are mobile phones annoying?," *Behaviour & Information Technology*, vol. 23, no. 1, pp. 33–41, 2004.

[100] D. Mook, "In defense of external invalidity.," *American psychologist*, vol. 38, no. 4, pp. 379–387, 1983.

[101] M. Moshagen and M. Thielsch, "Facets of visual aesthetics," *International Journal of Human-Computer Studies*, vol. 68, no. 10, pp. 689–709, 2010.

[102] F. Mueller, S. Agamanolis, and R. Picard, "Exertion interfaces: sports over a distance for social bonding and fun," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 561–568, New York, NY, USA: ACM, 2003.

[103] T. Munzner, "Process and pitfalls in writing information visualization research papers," in *Information Visualization: Human-Centered Issues and Perspectives*, (A. Kerren, J. Stasko, J.-D. Fekete, and C. North, eds.), pp. 134–153, Berlin: Springer, 2008.

[104] C. Nass, Y. Moon, B. J. Fogg, B. Reeves, and C. Dryer, "Can computer personalities be human personalities?," in *Conference Companion on Human Factors in Computing Systems*, pp. 228–229, New York, NY, USA: ACM, 1995.

[105] W. Newman and A. Taylor, "Towards a methodology employing critical parameters to deliver performance improvements in interactive systems," in *Proceedings IFIP TC13 Seventh International Conference on Human-Computer Interaction*, pp. 605–612, Amsterdam: IOS press, 1999.

[106] W. Newman, "A preliminary analysis of the products of HCI research, using pro forma abstracts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 278–284, New York, NY, USA: ACM, 1994.

[107] R. Nickerson, "Null hypothesis significance testing: a review of an old and continuing controversy.," *Psychological methods*, vol. 5, no. 2, p. 241, 2000.

[108] J. Nielsen, *Usability engineering*. Boston, MA: AP Professional, 1993.

[109] K. O'Hara and A. Sellen, "A comparison of reading paper and on-line documents," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 335–342, New York, NY: ACM Press, 1997.

[110] J. S. Olson, G. M. Olson, M. Storrøsten, and M. Carter, "Groupwork close up: a comparison of the group design process with and without a simple group editor," *ACM Transactions on Information Systems*, vol. 11, no. 4, pp. 321–348, Oct 1993.

[111] A. Oulasvirta, "Fielding Usability Evaluation," *IEEE Pervasive Computing*, vol. 11, no. 4, pp. 60–67, 2012.

[112] A. Oulasvirta, M. Wahlström, and K. Anders Ericsson, "What does it mean to be good at using a mobile device? An investigation of three levels of experience and skill," *International Journal of Human-computer Studies*, vol. 69, no. 3, pp. 155–169, 2011.

[113] T. Paek, S. Dumais, and R. Logan, "WaveLens: a new view onto Internet search results," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 727–734, New York, NY, USA: ACM, 2004.

[114] G. Paolacci, J. Chandler, and P. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010.

[115] D. Parnas and P. Clements, "A rational design process: How and why to fake it," *IEEE Transactions on Software Engineering*, no. 2, pp. 251–257, 1986.

[116] A. Pirhonen, S. Brewster, and C. Holguin, "Gestural and audio metaphors as a means of control for mobile devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 291–298, New York, NY, USA: ACM, 2002.

[117] J. Platt, "Strong inference," *Science*, vol. 146, no. 3642, pp. 347–353, 1964.

[118] T. L. Roberts and T. P. Moran, "The evaluation of text editors: methodology and empirical results.," *Communications of the ACM*, vol. 26, no. 4, pp. 265–283, Apr. 1983.

[119] C. Robson, *Real world research: A resource for social scientists and practitioner-researchers*. Oxford, UK: Blackwell, 2002.

[120] R. Rosenthal and R. Rosnow, *The volunteer subject*. New York, NY: John Wiley & Sons, 1975.

[121] R. Rosenthal and R. Rosnow, *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge, UK: Cambridge University Press, 1985.

[122] R. Rosenthal and R. Rosnow, *Essentials of behavioral research: Methods and data analysis*. New York, NY: McGraw-Hill, 1991.

[123] P. M. Sanderson and C. Fisher, "Exploratory sequential data analysis: foundations," *Human-computer Interaction*, vol. 9, no. 4, pp. 251–317, Sep 1994.

[124] J. Sauro and J. R. Lewis, "Correlations among prototypical usability metrics: evidence for the construct of usability," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1609–1618, New York, NY, USA: ACM, 2009.

[125] A. Sears and B. Shneiderman, "Split menus: effectively using selection frequency to organize menus," *ACM Transactions on Computer-Human Interaction*, vol. 1, no. 1, pp. 27–51, Mar 1994.

[126] D. Sears, "College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature.," *Journal of Personality and Social Psychology*, vol. 51, no. 3, p. 515, 1986.

[127] W. R. Shadish, T. Cook, and D. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company, 2002.

[128] B. Shneiderman, *Designing the user interface*. Reading, MA: Addison-Wesley, 1987.

[129] B. Shneiderman and C. Plaisant, "Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies," in *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pp. 1–7, ACM, 2006.

[130] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, vol. 22, no. 11, pp. 1359–1366, 2011.

[131] D. Sjøberg, J. Hannay, O. Hansen, V. Kampenes, A. Karahasanovic, N. Liborg, and A. Rekdal, "A survey of controlled experiments in software engineering," *IEEE Transactions on Software Engineering*, vol. 31, no. 9, pp. 733–753, 2005.

[132] D. H. Sonnenwald, M. C. Whitton, and K. L. Maglaughlin, "Evaluating a scientific collaboratory: Results of a controlled experiment," *ACM Transactions on Computer-Human Interaction*, vol. 10, no. 2, pp. 150–176, Jun 2003.

[133] R. Soukoreff and I. MacKenzie, "Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI," *International Journal of Human-Computer Studies*, vol. 61, no. 6, pp. 751–789, 2004.

[134] W. Strunk and E. White, *The elements of style*. New York, NY: The Macmillan Company, 1959.

[135] W. Tichy, "Should computer scientists experiment more?," *Computer*, vol. 31, no. 5, pp. 32–40, 1998.

[136] E. Tufte, *Beautiful evidence*. Cheshire, CT: Graphics Press, 2006.

[137] T. Tullis and W. Albert, *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. San Francisco, CA: Morgan Kaufmann, 2008.

[138] E. Webb, D. Campbell, R. Schwartz, L. Sechrest, and J. Grove, *Nonreactive measures in the social sciences*. Boston: Houghton Mifflin, 1981.

[139] R. Weber, "Editor's comments: the rhetoric of positivism versus interpretivism: a personal view," *MIS quarterly*, vol. 28, no. 1, pp. iii–xii, 2004.

[140] S. Whittaker, L. Terveen, and B. Nardi, "Let's stop pushing the envelope and start addressing it: a reference task agenda for HCI," *Human–Computer Interaction*, vol. 15, no. 2-3, pp. 75–106, 2000.

[141] L. Wilkinson and T. T. F. on Statistical Inference, "Statistical methods in psychology journals: Guidelines and explanations.," *American psychologist*, vol. 54, no. 8, p. 594, 1999.

[142] D. Willer and H. Walker, *Building experiments: Testing social theory*. Stanford, CA: Stanford University Press, 2007.

[143] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 143–146, New York, NY, USA: ACM, 2011.

[144] J. Yi, Y. ah Kang, J. Stasko, and J. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.

[145] R. Yin, *Case study research: Design and methods*. Thousand Oaks, CA: Sage Publications, third ed., 2003.

[146] M. Zelkowitz and D. Wallace, "Experimental models for validating technology," *Computer*, vol. 31, no. 5, pp. 23–31, 1998.

[147] S. Zhai, S. Conversy, M. Beaudouin-Lafon, and Y. Guiard, "Human on-line response to target expansion," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 177–184, New York, NY, USA: ACM, 2003.