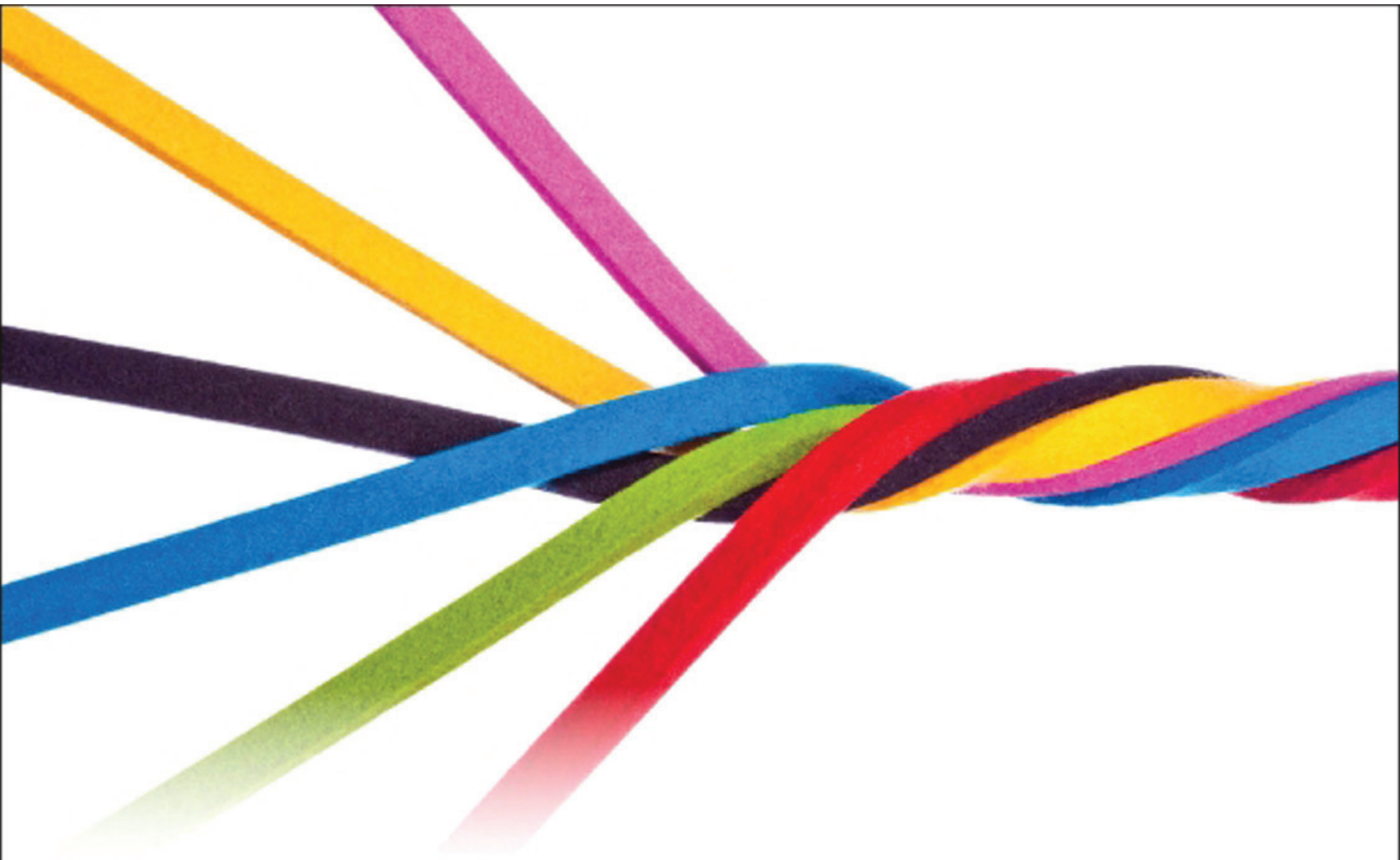

Human-Computer Interaction

An Empirical Research Perspective

MORGAN KAUFMANN

I. Scott MacKenzie

Human-Computer Interaction

CHAPTER

Scientific Foundations 4

In the last chapter, we examined a variety of interaction topics in HCI. By and large, the research methodology for studying these topics is empirical and scientific. Ideas are conceived, developed, and implemented and then framed as hypotheses that are tested in experiments. This chapter presents the enabling features of this methodology. Our goal is to establish the what, why, and how of research, with a focus on research that is both empirical and experimental. While much of the discussion is general, the examples are directed at HCI. We begin with the terminology surrounding research and empirical research.

4.1 What is research?

Research means different things to different people. "Being a researcher" or "conducting research" carries a certain elevated status in universities, colleges, and corporations. Consequently, the term research is bantered around in a myriad of situations. Often, the word is used simply to add weight to an assertion ("Our research shows that …"). While writing an early draft of this chapter, a television ad for an Internet service provider was airing in southern Ontario. The ad proclaimed, "Independent research proves [name_of_product] is the fastest and most reliable—period."[1] One might wonder about the nature of the research, or of the independence and impartiality of the work. Of course, forwarding assertions to promote facts, observations, hypotheses, and the like is often the goal. But what is research? Surely, it is more than just a word to add force to a statement or opinion. To rise above conjecture, we demand evidence—evidence meeting a standard of credibility such that the statement is beyond dispute. Providing such credibility is the goal of research.

Returning to the word itself, research has at least three definitions. First, conducting research can be an exercise as simple as *careful or diligent search*.[2] So carefully searching one's garden to find and remove weeds meets one standard of

[1] Advertisement by Rogers Communications Inc. airing on television in southern Ontario during the winter of 2008/2009.
[2] www.merriam-webster.com.

conducting research. Or perhaps one undertakes a search on a computer to locate all files modified on a certain date. That's research. It's not the stuff of MSc or PhD theses, but it meets one definition of research.

The second definition of research is *collecting information about a particular subject*. So surveying voters to collect information on political opinions is conducting research. In HCI we might observe people interacting with an interface and collect information about their interactions, such as the number of times they consulted the manual, clicked the wrong button, retried an operation, or uttered an expletive. That's research.

The third definition is more elaborate: *research is investigation or experimentation aimed at the discovery and interpretation of facts and revision of accepted theories or laws in light of new facts.*

In this definition we find several key elements of research that motivate discussions in this book. We find the idea of *experimentation*. Conducting experiments is a central activity in a lot of HCI research. I will say more about this in the next chapter. In HCI research, an experiment is sometimes called a *user study*. The methodology is sometimes formal, sometimes ad hoc. A formal and standardized methodology is generally preferred because it brings consistency to a body of work and facilitates the review and comparison of research from different studies. One objective of this book is to promote the use of a consistent methodology for experimental research in HCI.

To be fair, the title of this book changed a few times on the way to press. Is the book about *experimental research*? Well, yes, a lot of it is, but there are important forms of HCI research that are non-experimental. So as not to exclude these, the focus shifted to *empirical research*, a broader term that encompasses both experimental and non-experimental methodologies. Among the latter is building and testing models of interaction, which we examine formally in Chapter 7.

Returning to research, the third definition speaks of *facts*. Facts are the building blocks of evidence, and it is evidence we seek in experimental research. For example, we might observe that a user committed three errors while entering a command with an interface. That's a fact. Of course, context is important. Did the user have prior experience with the interface, or with similar interfaces? Was the user a child or a computer expert? Perhaps we observed and counted the errors committed by a group of users while interacting with two different interfaces over a period of time. If they committed 15 percent more errors with one interface than with the other, the facts are more compelling (but, again, context is important). Collectively, the facts form an outward sign leading to evidence—evidence that one interface is better, or less error prone, than the other. Evidence testing is presented in more detail in Chapter 6, Hypothesis Testing. Note that *prove* or *proof* is not used here. In HCI research we don't prove things; we gather facts and formulate and test evidence.

The third definition mentions *theories* and *laws*. Theory has two common meanings. In the sense of Darwin's *theory of evolution* or Einstein's *theory of relativity*, the term theory is synonymous with *hypothesis*. In fact, one definition of theory is simply "a hypothesis assumed for the sake of argument or investigation." Of course,

through experimentation, these theories advanced beyond argument and investigation. The stringent demands of scientific inquiry confirmed the hypotheses of these great scientists. When confirmed through research, a theory becomes a *scientifically accepted body of principles that explain phenomena*.

A *law* is different from a theory. A law is more specific, more constraining, more formal, more binding. In the most exacting terms, a law is a relationship or phenomenon that is "invariable under given conditions." Because variability is germane to human behavior, laws are of questionable relevance to HCI. Of course, HCI has laws. Take HCI's best-known law as an example. *Fitts' law* refers to a body of work, originally in human motor behavior (Fitts, 1954), but now widely used in HCI. Fitts' work pertained to rapid-aimed movements, such as rapidly moving a cursor to an object and selecting it in a graphical user interface. Fitts himself never proposed a law. He proposed a model of human motor behavior. And by all accounts, that's what Fitts' law is—a model, a behavioral, descriptive, and predictive model. It includes equations and such for predicting the time to do point-select tasks. It is a law only in that other researchers took up the label as a celebration of the generality and importance of Fitts' seminal work. We should all be so lucky. Fitts' law is presented in more detail in Chapter 7.

Research, according to the third definition, involves *discovery*, *interpretation*, and *revision*. Discovery is obvious enough. That's what we do—look for, or discover, things that are new and useful. Perhaps the discovery is a new style of interface or a new interaction technique. Interpretation and revision are central to research. Research does not proceed in a vacuum. Today's research builds on what is already known or assumed. We interpret what is known; we revise and extend through discovery.

There are additional characteristics of research that are not encompassed in the dictionary definitions. Let's examine a few of these.

4.1.1 Research must be published

Publication is the final step in research. It is also an essential step. Never has this rung as true as in the edict *publish or perish*. Researchers, particularly in academia, must publish. A weak or insufficient list of publications might spell disappointment when applying for research funds or for a tenure-track professorship at a university. Consequently, developing the skill to publish begins as a graduate student and continues throughout one's career as a researcher, whether in academia or industry. The details and challenges in writing research papers are elaborated in Chapter 8.

Publishing is crucial, and for good reason. Until it is published, the knowledge gained through research cannot achieve its critical purpose—to extend, refine, or revise the existing body of knowledge in the field. This is so important that publication bearing a high standard of scrutiny is required. Not just any publication, but publication in archived peer-reviewed journals or conference proceedings. Research results are "written up," submitted, and reviewed for their integrity, relevance, and contribution. The review is by peers—other researchers doing similar work. Are

the results novel and useful? Does the evidence support the conclusions? Is there a contribution to the field? Does the methodology meet the expected standards for research? If these questions are satisfactorily answered, the work has a good chance of acceptance and publication. Congratulations. In the end, the work is published and archived. *Archived* implies the work is added to the collection of related work accessible to other researchers throughout the world. This is the "existing body of knowledge" referred to earlier. The final step is complete.

Research results are sometimes developed into bona fide inventions. If an individual or a company wishes to profit from their invention, then patenting is an option. The invention is disclosed in a patent application, which also describes previous related work (prior art), how the invention addresses a need, and the best mode of implementation. If the application is successful, the patent is granted and the inventor or company thereafter owns the rights to the invention. If another company wishes to use the invention for commercial purpose, they must enter into a license agreement with the patent holder. This side note is included only to make a small point: a patent is a publication. By patenting, the individual or company is not only retaining ownership of the invention but is also making it public through publication of the patent. Thus, patents meet the must-publish criterion for research.

4.1.2 Citations, references, impact

Imagine the World Wide Web without hyperlinks. Web pages would live in isolation, without connections between them. Hyperlinks provide the essential pathways that connect web pages to other web pages, thus providing structure and cohesion to a topic or theme. Similarly, it is hard to imagine the world's body of published research without *citations* and *references*. Citations, like hyperlinks, connect research papers to other research papers. Through citations, a body of research takes shape. The insights and lessons of early research inform and guide later research. The citation itself is just an abbreviated tag that appears in the body of a paper, for example, "… as noted in earlier research (Smith and Jones, 2003)" or "… as confirmed by Green et al. [5]." These two examples are formatted differently and follow the requirements of the conference or journal. The citation is expanded into a full bibliographic entry in the reference list at the end of the paper. Formatting of citations and references is discussed in Chapter 8.

Citations serve many purposes, including supporting intellectual honesty. By citing previous work, researchers acknowledge that their ideas continue, extend, or refine those in earlier research. Citations are also important to back up assertions that are otherwise questionable, for example, "the number of tablet computer users worldwide now exceeds two billion [9]." In the Results section of a research paper, citations are used to compare the current results with those from earlier research, for example, "the mean time to formulate a search query was about 15 percent less than the time reported by Smith and Jones [5]."

Figure 4.1 provides a schematic of a collection of research papers. Citations are shown as arrows. It incorporates a timeline, so all arrows point to the left, to earlier

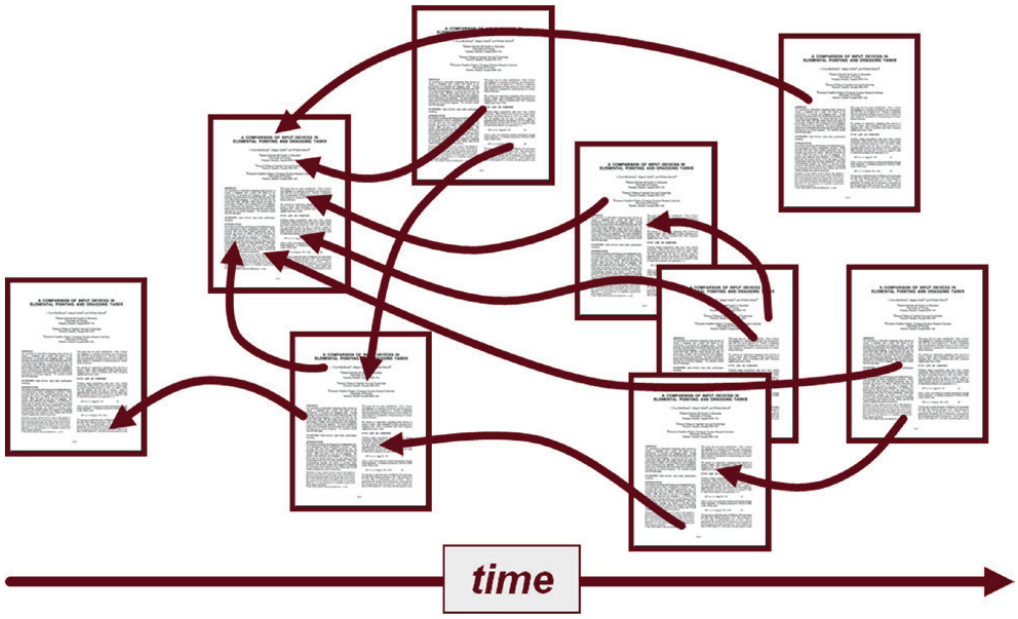

FIGURE 4.1

A collection of research papers with citations to earlier papers.

papers. One of the papers seems to have quite a few citations to it. The number of citations to a research paper is a measure of the paper's *impact*. If many researchers cite a single paper, there is a good chance the work described in the cited paper is both of high quality and significant to the field. This point is often echoed in academic circles: "The only objective and transparent metric that is highly correlated with the quality of a paper is the number of citations."[3] Interestingly enough, citation counts are only recently easily available. Before services like Google Scholar emerged, citation counts were difficult to obtain.

Since citation counts are available for individual papers, they are also easy to compile for individual researchers. Thus, impact can be assessed for researchers as well as for papers. The most accepted single measure of the impact of a researcher's publication record is the *H-index*. If a researcher's publications are ordered by the number of citations to each paper, the H-index is the point where the rank equals the number of citations. In other words, a researcher with H-index = n has n publications each with n or more citations. Physicist J. Hirsch first proposed the H-index in 2005 (Hirsch, 2005). H-index quantifies in a single number both research productivity (number of publications) and overall impact of a body of work (number of citations). Some of the strengths and weaknesses of the H-index, as a measure of impact, are elaborated elsewhere (MacKenzie, 2009a).

[3]Dianne Murray, General Editor, *Interacting with Computers*. Posted to chi-announcements@acm.org on Oct 8, 2008.

4.1.3 Research must be reproducible

Research that cannot be replicated is useless. Achieving an expected standard of reproducibility, or repeatability, is therefore crucial. This is one reason for advancing a standardized methodology: it enforces a process for conducting and writing about the research that ensures sufficient detail is included to allow the results to be replicated. If skilled researchers care to test the claims, they will find sufficient guidance in the methodology to reproduce, or replicate, the original research. This is an essential characteristic of research.

Many great advances in science and research pertain to methodology. A significant contribution by Louis Pasteur (1822–1895), for example, was his use of a consistent methodology for his research in microbiology (Day and Gastel, 2006, pp. 8–9). Pasteur's experimental findings on germs and diseases were, at the time, controversial. As Pasteur realized, the best way to fend off skepticism was to empower critics—other scientists—to see for themselves. Thus, he adopted a methodology that included a standardized and meticulous description of the materials and procedure. This allowed his experiments and findings to be replicated. A researcher questioning a result could redo the experiment and therefore verify or refute the result. This was a crucial advance in science. Today, reviewers of manuscripts submitted for publication are often asked to critique the work on this very point: "Is the work replicable?" "No" spells certain rejection.

One of the most cited papers in publishing history is a method paper. Lowry et al.'s, 1951 paper "Protein Measurement With the Folin Phenol Reagent" has garnered in excess of 200,000 citations (Lowry, Rosenbrough, Farr, and Randall, 1951).[4] The paper describes a method for measuring proteins in fluids. In style, the paper reads much like a recipe. The method is easy to read, easy to follow, and, importantly, easy to reproduce.

4.1.4 Research versus engineering versus design

There are many ways to distinguish research from engineering and design. Researchers often work closely with engineers and designers, but the skills and contributions each brings are different. Engineers and designers are in the business of building things. They create products that strive to bring together the best in *form* (design emphasis) and *function* (engineering emphasis). One can imagine that there is certain tension, even a trade-off, between form and function. Finding the right balance is key. However, sometimes the balance tips one way or the other. When this occurs, the result is a product or a feature that achieves one (form or function) at the expense of the other. An example is shown in Figure 4.2a. The image shows part of a notebook computer, manufactured by a well-known computer company. By most accounts, it is a typical notebook computer. The image shows part of the keyboard and the built-in pointing device, a touchpad. The touchpad design (or is it engineering?) is interesting. It is seamlessly embedded in the system chassis.

[4] See http://scholar.google.com.

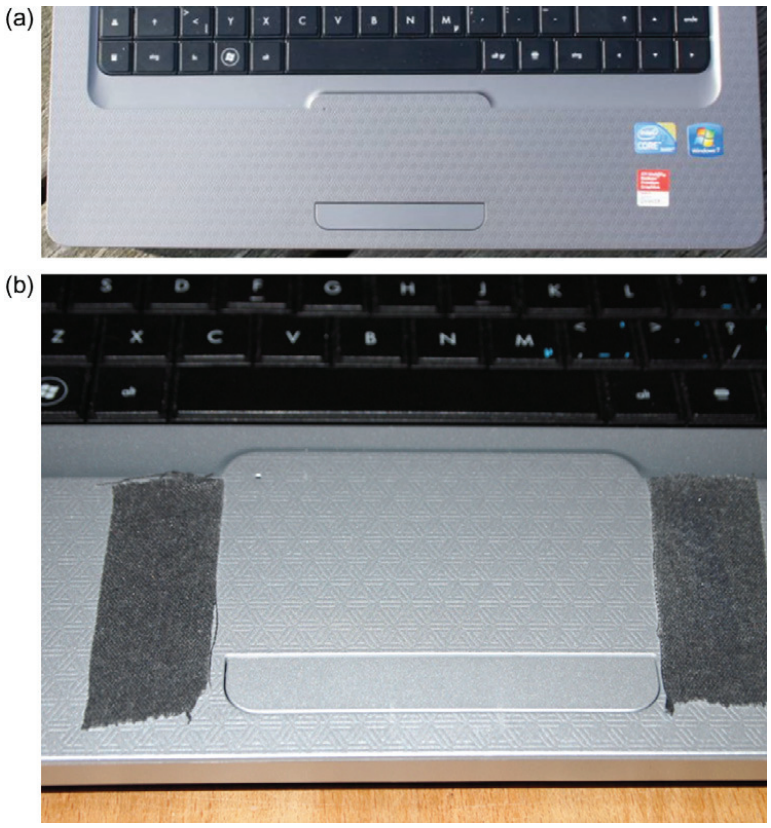

FIGURE 4.2

Form trumping function: (a) Notebook computer. (b) Duct tape provides tactile feedback indicating the edge of the touchpad.

The look is elegant—smooth, shiny, metallic. But something is wrong. Because the mounting is seamless and smooth, tactile feedback at the sides of the touchpad is missing. While positioning a cursor, the user has no sense of when his or her finger reaches the edge of the touchpad, except by observing that the cursor ceases to move. This is an example of form trumping function. One user's solution is shown in Figure 4.2b. Duct tape added on each side of the touchpad provides the all-important tactile feedback.[5]

Engineers and designers work in the world of products. The focus is on designing complete systems or products. Research is different. Research tends to be narrowly focused. Small ideas are conceived of, prototyped, tested, then advanced or discarded. New ideas build on previous ideas and, sooner or later, good ideas are refined into the building blocks—the materials and processes—that find their way into products. But research questions are generally small in scope. Research tends to be incremental, not monumental.

[5]For an amusing example of function trumping form, visit Google Images using "Rube Goldberg simple alarm clock."

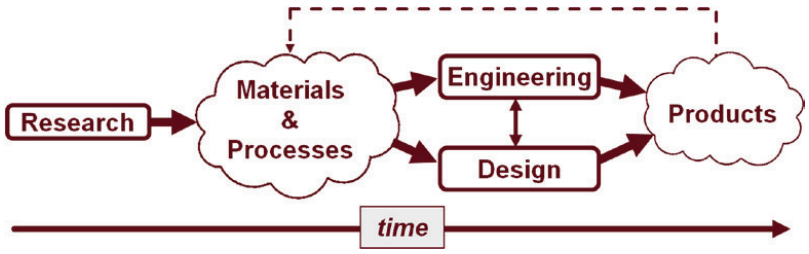

FIGURE 4.3

Timeline for research, engineering, and design.

Engineers and designers also work with prototypes, but the prototype is used to assess alternatives at a relatively late stage: as part of product development. A researcher's prototype is an early mock-up of an idea, and is unlikely to directly appear in a product. Yet the idea of using prototypes to inform or assess is remarkably similar, whether for research or for product development. The following characterization by Tim Brown (CEO of design firm IDEO) is directed at designers, but is well aligned with the use of prototypes for research:

> *Prototypes should command only as much time, effort, and investment as are needed to generate useful feedback and evolve an idea. The more "finished" a prototype seems, the less likely its creators will be to pay attention to and profit from feedback. The goal of prototyping isn't to finish. It is to learn about the strengths and weaknesses of the idea and to identify new directions that further prototypes might take (Brown, 2008, p. 3).*

One facet of research that differentiates it from engineering and design is the timeline. Research precedes engineering and design. Furthermore, the march forward for research is at a slower pace, without the shackles of deadlines. Figure 4.3 shows the timeline for research, engineering, and design. Products are the stuff of deadlines. Designers and engineers work within the corporate world, developing products that sell, and hopefully sell well. The raw materials for engineers and designers are materials and processes that already exist (dashed line in Figure 4.3) or emerge through research.

The computer mouse is a good example. It is a hugely successful product that, in many ways, defines a generation of computing, post 1981, when the Xerox Star was introduced. But in the 1960s the mouse was just an idea. As a prototype it worked well as an input controller to maneuver a tracking symbol on a graphics display. Engelbart's invention (English et al., 1967) took nearly 20 years to be engineered and designed into a successful product.

Similar stories are heard today. Apple Computer Inc., long known as a leader in innovation, is always building a better mousetrap. An example is the *iPhone*, introduced in June, 2007. And, evidently, the world has beaten a path to Apple's door.[6] Notably, "with the iPhone, Apple successfully brought together decades

[6]The entire quotation is "Build a better mousetrap and the world will beat a path to your door" and is attributed to American essayist Ralph Waldo Emerson (1803–1882).

of research" (Selker, 2008). Many of the raw materials of this successful product came by way of low-level research, undertaken well before Apple's engineers and designers set forth on their successfully journey. Among the iPhone's interaction novelties is a two-finger *pinch* gesture for zooming in and out. New? Perhaps, but Apple's engineers and designers no doubt were guided or inspired by research that came before them. For example, multi-touch gestures date back to at least the 1980s (Buxton, Hill, and Rowley, 1985; Hauptmann, 1989). What about changing the aspect ratio of the display when the device is tilted? New? Perhaps not. Tilt, as an interaction technique for user interfaces, dates back to the 1990s (B. Harrison et al., 1998; Hinckley et al., 2000; Rekimoto, 1996). These are just two examples of research ideas that, taken alone, are small scale. While engineers and designers strive to build better systems or products, in the broadest sense, researchers provide the raw materials and processes engineers and designers work with: stronger steel for bridges, a better mouse for pointing, a better algorithm for a search engine, a more natural touch interface for mobile phones.

4.2 What is empirical research?

By prefixing research with *empirical*, some powerful new ideas are added. According to one definition, empirical means *originating in or based on observation or experience*. Simple enough. Another definition holds that empirical means *relying on experience or observation alone, often without due regard for system and theory*. This is interesting. These words suggest researchers should be guided by direct observations and experiences about phenomena, without prejudice to, or even consideration of, existing theories. This powerful idea is a guiding principle in science—not to be blinded by preconceptions. Here's an example. Prior to the 15th century, there was a prevailing *system* or *theory* that celestial bodies revolved around the earth. The Polish scientist Nicolas Copernicus (1473–1543) found evidence to the contrary. His work was empirical. It was based on observation without bias toward, influence by, or due regard to, existing theory. He observed, he collected data, he looked for patterns and relationships in the data, and he found evidence within the data that cut across contemporary thinking. His empirical evidence led to one of the great achievements in modern science—a heliocentric cosmology that placed the sun, rather than the earth, at the center of the solar system. Now that's a nice *discovery* (see the third definition of research at the beginning of this chapter). In HCI and other fields of research, discoveries are usually more modest.

By another definition, empirical means *capable of being verified or disproved by observation or experiment*. These are strong words. An HCI research initiative is framed by hypotheses—assertions about the merits of an interface or an interaction technique. The assertions must be sufficiently clear and narrow to enable verification or disproval by gathering and testing evidence. This means using language in an assertion that speaks directly to empirical, observable, quantifiable aspects of the interaction. I will expand on this later in this chapter in the discussion on research questions.

4.3 Research methods

There are three common approaches, or methods, for conducting research in HCI and other disciplines in the natural and social sciences: the *observational method*, the *experimental method*, and the *correlational method*. All three are empirical as they are based on observation or experience. But there are differences and these follow from the objectives of the research and from the expertise and style of the researcher. Let's examine each method.

4.3.1 Observational method

Observation is the starting point for this method. In conducting empirical research in HCI, it is essential to observe humans interacting with computers or computer-embedded technology of some sort. The observational method encompasses a collection of common techniques used in HCI research. These include interviews, field investigations, contextual inquiries, case studies, field studies, focus groups, think aloud protocols, storytelling, walkthroughs, cultural probes, and so on. The approach tends to be qualitative rather than quantitative. As a result, observational methods achieve *relevance* while sacrificing *precision* (Sheskin, 2011, p. 76). Behaviors are studied by directly observing phenomena in a natural setting, as opposed to crafting constrained behaviors in an artificial laboratory setting. Real world phenomena are high in relevance, but lack the precision available in controlled laboratory experiments.

Observational methods are generally concerned with discovering and explaining the reasons underlying human behavior. In HCI, this is the *why* or *how* of the interaction, as opposed to the *what*, *where*, or *when*. The methods focus on human thought, feeling, attitude, emotion, passion, sensation, reflection, expression, sentiment, opinion, mood, outlook, manner, style, approach, strategy, and so on. These human qualities can be studied through observational methods, but they are difficult to measure. The observations are more likely to involve note-taking, photographs, videos, or audio recordings rather than measurement. Measurements, if gathered, tend to use categorical data or simple counts of phenomena. Put another way, observational methods tend to examine and record the quality of interaction rather than quantifiable human performance.

4.3.2 Experimental method

With the experimental method (also called the *scientific method*), knowledge is acquired through controlled experiments conducted in laboratory settings. Acquiring knowledge may imply gathering new knowledge, but it may also mean studying existing knowledge for the purpose of verifying, refuting, correcting, integrating, or extending. In the relevance-precision dichotomy, it is clear where controlled experiments lie. Since the tasks are artificial and occur in a controlled laboratory setting, relevance is diminished. However, the control inherent in the

methodology brings precision, since extraneous factors—the diversity and chaos of the real world—are reduced or eliminated.

A controlled experiment requires at least two variables: a *manipulated variable* and a *response variable*. In HCI, the manipulated variable is typically a property of an interface or interaction technique that is presented to participants in different configurations. Manipulating the variable simply refers to systematically exposing participants to different configurations of the interface or interaction technique. To qualify as a controlled experiment, at least two configurations are required. Thus, comparison is germane to the experimental method. This point deserves further elaboration. In HCI, we often hear of a system or design undergoing a "usability evaluation" or "user testing." Although these terms often have different meanings in different contexts, such evaluations or tests generally do not follow the experimental method. The reason is simple: there is no manipulated variable. This is mentioned only to distinguish a usability evaluation from a *user study*. Undertaking a user study typically implies conducting a controlled experiment where different configurations of a system are tested and compared. A "usability evaluation," on the other hand, usually involves assessing a single user interface for strengths and weaknesses. The evaluation might qualify as research ("collecting information about a particular subject"), but it is not experimental research. I will return to this point shortly. A manipulated variable is also called an *independent variable* or *factor*.

A response variable is a property of human behavior that is observable, quantifiable, and measurable. The most common response variable is time, often called *task completion time* or some variation thereof. Given a task, how long do participants take to do the task under each of the configurations tested? There are, of course, a multitude of other behaviors that qualify as response variables. Which ones are used depend on the characteristics of the interface or interaction technique studied in the research. A response variable is also called a *dependent variable*. Independent variables and dependent variables are explored in greater detail in Chapter 5.

HCI experiments involve humans, so the methodology employed is borrowed from experimental psychology, a field with a long history of research involving humans. In a sense, HCI is the beneficiary of this more mature field. The circumstances manipulated in a psychology experiment are often quite different from those manipulated in an HCI experiment, however. HCI is narrowly focused on the interaction between humans and computing technology, while experimental psychology covers a much broader range of the human experience.

It is naïve to think we can simply choose to focus on the experimental method and ignore qualities of interaction that are outside the scope of the experimental procedure. A full and proper user study—an experiment with human participants—involves more than just measuring and analyzing human performance. We engage observational methods by soliciting comments, thoughts, and opinions from participants. Even though a task may be performed quickly and with little or no error, if participants experience fatigue, frustration, discomfort, or another *quality* of interaction, we want to know about it. These qualities of interaction may not appear in the numbers, but they cannot be ignored.

One final point about the experimental method deserves mention. A controlled experiment, if designed and conducted properly, often allows a powerful form of conclusion to be drawn from the data and analyses. The relationship between the independent variable and the dependent variable is one of *cause and effect*; that is, the manipulations in the interface or interaction techniques are said to have *caused* the observed differences in the response variable. This point is elaborated in greater detail shortly. Cause-and-effect conclusions are not possible in research using the observational method or the correlational method.

4.3.3 Correlational method

The correlational method involves looking for relationships between variables. For example, a researcher might be interested in knowing if users' privacy settings in a social networking application are related to their personality, IQ, level of education, employment status, age, gender, income, and so on. Data are collected on each item (privacy settings, personality, etc.) and then relationships are examined. For example, it might be apparent in the data that users with certain personality traits tend to use more stringent privacy settings than users with other personality traits.

The correlational method is characterized by quantification since the magnitude of variables must be ascertained (e.g., age, income, number of privacy settings). For nominal-scale variables, categories are established (e.g., personality type, gender). The data may be collected through a variety of methods, such as observation, interviews, on-line surveys, questionnaires, or measurement. Correlational methods often accompany experimental methods, if questionnaires are included in the experimental procedure. Do the measurements on response variables suggest relationships by gender, by age, by level of experience, and so on?

Correlational methods provide a balance between relevance and precision. Since the data were not collected in a controlled setting, precision is sacrificed. However, data collected using informal techniques, such as interviews, bring relevance—a connection to real-life experiences. Finally, the data obtained using correlational methods are circumstantial, not causal. I will return to this point shortly.

This book is primarily directed at the experimental method for HCI research. However, it is clear in the discussions above that the experimental method will often include observational methods and correlational methods.

4.4 Observe and measure

Let's return to the foundation of empirical research: observation.

4.4.1 Observation

The starting point for empirical research in HCI is to observe humans interacting with computers. But how are observations made? There are two possibilities. Either

another human is the observer or an apparatus is the observer. A human observer is the experimenter or investigator, not the human interacting with the computer. Observation is the precursor to *measurement*, and if the investigator is the observer, then measurements are collected manually. This could involve using a log sheet or notebook to jot down the number of events of interest observed. Events of interest might include the number of times the user clicked a button or moved his or her hand from the keyboard to the mouse. It might involve observing users in a public space and counting those who are using mobile phones in a certain way, for example, while walking, while driving, or while paying for groceries at a checkout counter. The observations may be broken down by gender or some other attribute of interest.

Manual observation could also involve timing by hand the duration of activities, such as the time to type a phrase of text or the time to enter a search query. One can imagine the difficulty in manually gathering measurements as just described, not to mention the inaccuracy in the measurements. Nevertheless, manual timing is useful for preliminary testing, sometimes called *pilot testing*.

More often in empirical research, the task of observing is delegated to the apparatus—the computer. Of course, this is a challenge in some situations. As an example, if the interaction is with a digital sports watch or automated teller machine (ATM), it is not possible to embed data collection software in the apparatus. Even if the apparatus is a conventional desktop computer, some behaviors of interest are difficult to detect. For example, consider measuring the number of times the user's attention switches from the display to the keyboard while doing a task. The computer is not capable of detecting this behavior. In this case, perhaps an eye tracking apparatus or camera could be used, but that adds complexity to the experimental apparatus. Another example is clutching with a mouse—lifting and repositioning the device. The data transmitted from a mouse to a host computer do not include information on clutching, so a conventional host system is not capable of observing and recording this behavior. Again, some additional apparatus or sensing technology may be devised, but this complicates the apparatus. Or a human observer can be used. So depending on the behaviors of interest, some ingenuity might be required to build an apparatus and collect the appropriate measurements.

If the apparatus includes custom software implementing an interface or interaction technique, then it is usually straightforward to record events such as key presses, mouse movement, selections, finger touches, or finger swipes and the associated timestamps. These data are stored in a file for follow-up analyses.

4.4.2 Measurement scales

Observation alone is of limited value. Consider observations about rain and flowers. In some locales, there is ample rain but very few flowers in April. This is followed by less rain and a full-blown field of flowers in May. The observations may inspire anecdote (*April showers bring May flowers*), but a serious examination of patterns for rain and flowers requires measurement. In this case, an observer located in a garden would observe, measure, and record the amount of rain and

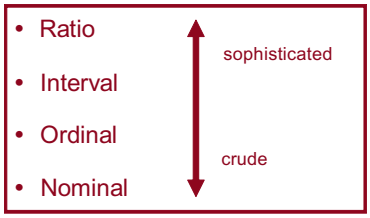

FIGURE 4.4

Scales of measurement: nominal, ordinal, interval, and ratio. Nominal measurements are considered simple, while ratio measurements are sophisticated.

the number of flowers in bloom. The measurements might be recorded each day during April and May, perhaps by several observers in several gardens. The measurements are collected, together with the means, tallied by month and analyzed for "significant differences" (see Chapter 6). With measurement, anecdotes turn to empirical evidence. The observer is now in a position to quantify the amount of rain and the number of flowers in bloom, separately for April and May. The added value of measurement is essential for science. In the words of engineer and physicist Lord Kelvin (1824–1907), after whom the Kelvin scale of temperature is named, "[Without measurement] your knowledge of it is of a meager and unsatisfactory kind."[7]

As elaborated in many textbooks on statistics, there are four scales of measurement: nominal, ordinal, interval, and ratio. Organizing this discussion by these four scales will help. Figure 4.4 shows the scales along a continuum with nominal scale measurements as the least sophisticated and ratio-scale measurements as the most sophisticated. This follows from the types of computations possible with each measurement, as elaborated below.

The nature, limitations, and abilities of each scale determine the sort of information and analyses possible in a research setting. Each is briefly defined below.

4.4.3 Nominal

A measurement on the nominal scale involves arbitrarily assigning a code to an attribute or a category. The measurement is so arbitrary that the code needn't be a number (although it could be). Examples are automobile license plate numbers, codes for postal zones, job classifications, military ranks, etc. Clearly, mathematical manipulations on nominal data are meaningless. It is nonsense, for example, to

[7]The exact and full quote, according to several online sources, is "When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge of it is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced it to the stage of science."

compute the mean of several license plate numbers. Nominal data identify mutually exclusive categories. Membership or exclusivity is meaningful, but little else. The only relationship that holds is equivalence, which exists between entities in the same class. Nominal data are also called *categorical data.*

If we are interested in knowing whether males and females differ in their use of mobile phones, we might begin our investigation by observing people and assigning each a code of "M" for male, "F" for female. Here, the attribute is gender and the code is M or F. If we are interested in handedness, we might observe the writing habits of users and assign codes of "LH" for left-handers and "RH" for right-handers. If we are interested in scrolling strategies, we might observe users interacting with a GUI application and categorize them according to their scrolling methods, for example as "MW" for mouse wheel, "CD" for clicking and dragging the scrollbar, or "KB" for keyboard.

Nominal data are often used with frequencies or counts—the number of occurrences of each attribute. In this case, our research is likely concerned with the difference in the counts between categories: "Are males or females more likely to …?", "Do left handers or right handers have more difficulty with …?", or "Are Mac or PC users more inclined to …?" Bear in mind that while the attribute is categorical, the count is a ratio-scale measurement (discussed shortly).

Here is an example of nominal scale attributes using real data. Attendees of an HCI research course were dispatched to several locations on a university campus. Their task was to observe, categorize, and count students walking between classes. Each student was categorized by gender (male, female) and by whether he or she was using a mobile phone (not using, using). The results are shown in Figure 4.5. A total of 1,527 students were observed. The split by gender was roughly equal (51.1% male, 48.9% female). By mobile phone usage, 13.1 percent of the students (200) were observed using their mobile phone while walking.

The research question in Figure 4.5 is a follows: are males or females more likely to use a mobile phone as they walk about a university campus? I will demonstrate how to answer this question in Chapter 6 on Hypothesis Testing.

Gender	Mobile Phone Usage		Total	%
	Not Using	Using		
Male	683	98	781	51.1%
Female	644	102	746	48.9%
Total	1327	200	1527	
%	86.9%	13.1%		

FIGURE 4.5

Two examples of nominal scale data: gender (male, female) and mobile phone usage (not using, using).

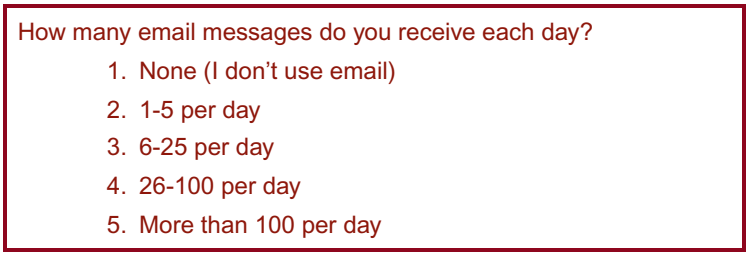

FIGURE 4.6

Example of a questionnaire item soliciting an ordinal response.

4.4.4 Ordinal data

Ordinal scale measurements provide an order or ranking to an attribute. The attribute can be any characteristic or circumstance of interest. For example, users might be asked to try three global positioning systems (GPS) for a period of time and then rank the systems by preference: first choice, second choice, third choice. Or users could be asked to consider properties of a mobile phone such as price, features, cool-appeal, and usability, and then order the features by personal importance. One user might choose usability (first), cool-appeal (second), price (third), and then features (fourth). The main limitation of ordinal data is that the interval is not intrinsically equal between successive points on the scale. In the example just cited, there is no innate sense of how much more important usability is over cool-appeal or whether the difference is greater or less than that between, for example, cool-appeal and price.

If we are interested in studying users' e-mail habits, we might use a questionnaire to collect data. Figure 4.6 gives an example of a questionnaire item soliciting ordinal data. There are five rankings according to the number of e-mail messages received per day. It is a matter of choice whether to solicit data in this manner or, in the alternative, to ask for an estimate of the number of e-mail messages received per day. It will depend on how the data are used and analyzed.

Ordinal data are slightly more sophisticated than nominal data since comparisons of *greater than* or *less than* are possible. However, it is not valid to compute the mean of ordinal data.

4.4.5 Interval data

Moving up in sophistication, interval data have equal distances between adjacent values. However, there is no absolute zero. The classic example of interval data is temperature measured on the Fahrenheit or Celsius scale. Unlike ordinal data, it is meaningful to compute the mean of interval data, for example, the mean mid-day temperature during the month of July. Ratios of interval data are not meaningful, however. For example, one cannot say that 20°C is twice as warm as 10°C.

In HCI, interval data are commonly used in questionnaires where a response on a linear scale is solicited. An example is a Likert Scale (see Figure 4.7), where verbal responses are given a numeric code. In the example, verbal responses are

Please indicate your level of agreement with the following statements.

	Strongly disagree	Mildly disagree	Neutral	Mildly agree	Strongly agree
It is safe to talk on a mobile phone while driving.	1	2	3	4	5
It is safe to read a text message on a mobile phone while driving.	1	2	3	4	5
It is safe to compose a text message on a mobile phone while driving.	1	2	3	4	5

FIGURE 4.7

A set of questionnaire items organized in a Likert Scale. The responses are examples of interval scale data.

symmetric about a neutral, central value with the gradations between responses more or less equal. It is this last quality—equal gradations between responses—that validates calculating the mean of the responses across multiple respondents.

There is some disagreement among researchers on the assumption of equal gradations between the items in Figure 4.7. Do respondents perceive the difference between, say, 1 and 2 (strongly disagree and mildly disagree) the same as the difference between, say, 2 and 3 (mildly disagree and neutral)? Attaching verbal tags to numbers is likely to bring qualitative and highly personal interpretations to the responses. There is evidence that respondents perceive items at the extremes of the scale as farther apart than items in the center (Kaptein, Nass, and Markopoulos, 2010). Nevertheless, the graduation between responses is much more similar here than between the five ordinal responses in Figure 4.6. One remedy for non-equal gradations in Likert-scale response items is simply to instruct respondents to interpret the items as equally spaced.

Examples of Likert Scale questionnaire items in HCI research papers are as follows: Bickmore and Picard, 2004; Dautenhahn et al., 2006; Garau et al., 2003; Guy, Ur, Ronen, Perer, and Jacovi, 2011; Wobbrock, Chau, and Myers, 2007.

4.4.6 Ratio data

Ratio-scale measurements are the most sophisticated of the four scales of measurement. Ratio data have an absolute zero and support a myriad of calculations to

summarize, compare, and test the data. Ratio data can be added, subtracted, multiplied, divided; means, standard deviations, and variances can be computed. In HCI, the most common ratio-scale measurement is time—the time to complete a task. But generally, all physical measurements are also ratio-scale, such as the distance or velocity of a cursor as it moves across a display, the force applied by a finger on a touchscreen, and so on. Many social variables are also ratio-scale, such as a user's age or years of computer experience.

Another common ratio-scale measurement is count (noted above). Often in HCI research, we count the number of occurrences of certain human activities, such as the number of button clicks, the number of corrective button clicks, the number of characters entered, the number of incorrect characters entered, the number of times an option is selected, the number of gaze shifts, the number of hand movements between the mouse and keyboard, the number of task retries, the number of words in a search query, etc. Although we tend to give time special attention, it too is a count—the number of seconds or minutes elapsed as an activity takes place. These are all ratio-scale measurements.

The expressive nature of a count is improved through *normalization*; that is, expressing the value as a count *per something*. So for example, knowing that a 10-word phrase was entered in 30 seconds is less revealing than knowing that the rate of entry was $10/0.5 = 20$ words per minute (wpm). The main benefit of normalizing counts is to improve comparisons. It is easy to compare 20 wpm for one method with 23 wpm for another method—the latter method is faster. It is much harder to compare 10 words entered in 30 seconds for one method with 14 words entered in 47 seconds for another method.

As another example, let's say two errors were committed while entering a 50-character phrase of text. Reporting the occurrence of two errors reveals very little, unless we also know the length of the phrase. Even so, comparisons with results from another study are difficult. (What if the other study used phrases of different lengths?) However, if the result is reported as a $2/50 = 4\%$ error rate, there is an immediate sense of the meaning, magnitude, and relevance of the human performance measured, and as convention has it, the other study likely reported error rates in much the same way. So where possible, normalize counts to make the measurements more meaningful and to facilitate comparisons.

An example in the literature is an experiment comparing five different text entry methods (Magerkurth and Stenzel, 2003). For speed, results were reported in "words per minute" (that's fine); however, for accuracy, results were reported as the number of errors committed. Novice participants, for example, committed 24 errors while using multi-tap (Magerkurth and Stenzel, 2003, Table 2). While this number is useful for comparing results within the experiment, it provides no insight as to how the results compare with those in related research. The results would be more enlightening if normalized for the amount of text entered and reported as an "error rate (%)," computed as the number of character errors divided by the total number of characters entered times 100.

4.5 Research questions

In HCI, we conduct experimental research to answer (and raise!) questions about a new or existing user interface or interaction technique. Often the questions pertain to the relationship between two variables, where one variable is a circumstance or condition that is manipulated (an interface property) and the other is an observed and measured behavioral response (task performance).

The notion of posing or answering questions seems simple enough, but this is tricky because of the human element. Unlike an algorithm operating on a data set, where the time to search, sort, or whatever is the same with each try, people exhibit variability in their actions. This is true both from person to person and for a single person repeating a task. The result is always different! This variability affects the confidence with which we can answer *research questions*. To gauge the confidence of our answers, we use statistical techniques, as presented in Chapter 6, Hypothesis Testing.

Research questions emerge from an inquisitive process. The researcher has an idea and wishes to see if it has merit. Initial thoughts are fluid and informal:

- Is it viable?
- Is it as good as or better than current practice?
- What are its strengths and weaknesses?
- Which of several alternatives is best?
- What are the human performance limits and capabilities?
- Does it work well for novices, for experts?
- How much practice is required to become proficient?

These questions are unquestionably relevant, since they capture a researcher's thinking at the early stages of a research project. However, the questions above suffer a serious deficiency: They are not testable. The goal, then, is to move forward from the loose and informal questions above to questions more suitable for empirical and experimental enquiry.

I'll use an example to show how this is done. Perhaps a researcher is interested in text entry on touchscreen phones. Texting is something people do a lot. The researcher is experienced with the Qwerty soft keyboard on touchscreen phones, but finds it error prone and slow. Having thought about the problem for a while, an idea emerges for a new technique for entering text. Perhaps it's a good idea. Perhaps it's really good, better than the basic Qwerty soft keyboard (QSK). Being motivated to do research in HCI, the researcher builds a prototype of the entry technique and fiddles with the implementation until it works fine. The researcher decides to undertake some experimental research to evaluate the idea. What are the research questions? Perhaps the following capture the researcher's thinking:

- Is the new technique any good?
- Is the new technique better than QSK?

- Is the new technique faster than QSK?
- Is the new technique faster than QSK after a bit of practice?
- Is the measured entry speed (in words per minute) higher for the new technique than for a QSK after one hour of use?

From top to bottom, the questions are progressively narrower and more focused. Expressions like "any good" or "better than," although well intentioned, are problematic for research. Remember observation and measurement? How does one measure "better than"? Farther down the list, the questions address qualities that are more easily observed and measured. Furthermore, since they are expressed across alternative designs, comparisons are possible. The last question speaks very specifically to entry speed measured in words per minute, to a comparison between two methods, and to a criterion for practice. This is a testable research question.

4.6 Internal validity and external validity

At this juncture we are in a position to consider two important properties of experimental research: *internal validity* and *external validity*. I'll use the research questions above to frame the discussion. Two of the questions appear in the plot in Figure 4.8. The *x*-axis is labeled Breadth of Question or, alternatively, External Validity. The *y*-axis is labeled Accuracy of Answer or, alternatively, Internal Validity.

The question

Is the new technique better than QSK?

is positioned as high in breadth (that's good!) yet answerable with low accuracy (that's bad!). As already noted, this question is not testable in an empirical sense. Attempts to answer it directly are fraught with problems, because we lack a methodology to observe and measure "better than" (even though finding better interfaces is the final goal).

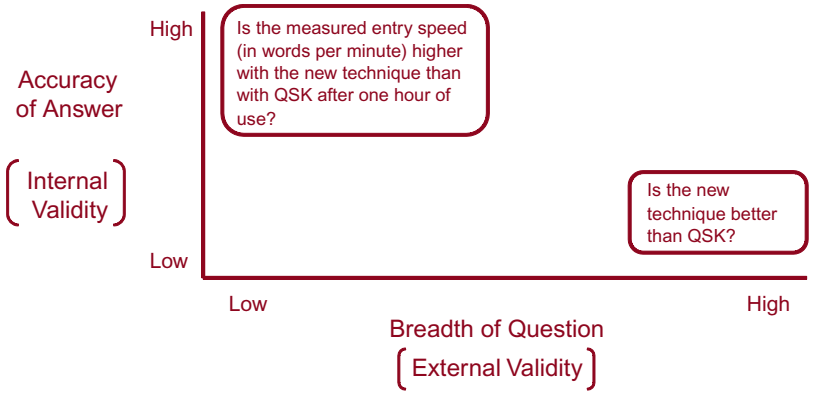

FIGURE 4.8

Graphical comparison of Internal Validity and External Validity.

The other, more detailed question

> Is the measured entry speed (in words per minute) higher with the new technique than with QSK after one hour of use?

is positioned as low in breadth (that's bad!) yet answerable with high accuracy (that's good!). The question is testable, which means we can craft a methodology to answer it through observation and measurement. Unfortunately, the narrow scope of the question brings different problems. Focusing on entry speed is fine, but what about other aspects of the interaction? What about accuracy, effort, comfort, cognitive load, user satisfaction, practical use of the technique, and so on? The question excludes consideration of these, hence the low breadth rating.

The alternative labels for the axes in Figure 4.8 are internal validity and external validity. In fact, the figure was designed to set up discussion on these important terms in experimental research.

Internal validity (definition) is the extent to which an effect observed is due to the test conditions. For the example, an *effect* is simply the difference in entry speed between the new technique and QSK. If we conduct an experiment to measure and compare the entry speed for the two techniques, we want confidence that the difference observed was actually due to inherent differences between the techniques. Internal validity captures this confidence. Perhaps the difference was due to something else, such as variability in the responses of the participants in the study. Humans differ. Some people are predisposed to be meticulous, while others are carefree, even reckless. Furthermore, human behavior—individually or between people—can change from one moment to the next, for no obvious reason. Were some participants tested early in the day, others late in the day? Were there any distractions, interruptions, or other environmental changes during testing? Suffice it to say that any source of variation beyond that due to the inherent properties of the test conditions tends to compromise internal validity. High internal validity means the effect observed really exists.

External validity (definition) is the extent to which experimental results are generalizable to other people and other situations. *Generalizable* clearly speaks to *breadth* in Figure 4.8. To the extent the research pursues broadly framed questions, the results tend to be broadly applicable. But there is more. Research results that apply to "other people" imply that the participants involved were representative of a larger intended population. If the experiment used 18- to 25-year-old computer literate college students, the results might generalize to middle-aged computer literate professionals. But they might not generalize to middle-aged people without computer experience. And they likely would not apply to the elderly, to children, or to users with certain disabilities. In experimental research, random sampling is important for generalizability; that is, the participants selected for testing were drawn at random from the desired population.

Generalizable to "other situations" means the experimental *environment* and *procedures* were representative of real world situations where the interface or

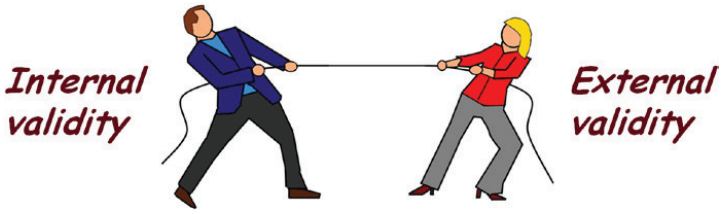

FIGURE 4.9

There is tension between internal validity and external validity. Improving one comes at the expense of the other.

(Sketch courtesy of Bartosz Bajer)

technique will be used. If the research studied the usability of a GPS system for taxi drivers or delivery personnel and the experiment was conducted in a quiet, secluded research lab, there may be a problem with external validity. Perhaps a different experimental environment should be considered. Research on text entry where participants enter predetermined text phrases with no punctuation symbols, no uppercase characters, and without any ability to correct mistakes, may have problem with external validity. Again, a different experimental procedure should be considered.

The scenarios above are overly dogmatic. Experiment design is an exercise in compromise. While speaking in the strictest terms about high internal validity and high external validity, in practice one is achieved at the expense of the other, as characterized in Figure 4.9.

To appreciate the tension between internal and external validity, two additional examples are presented. The first pertains to the experimental environment. Consider an experiment that compares two remote pointing devices for presentation systems. To improve external validity, the experimental environment mimics expected usage. Participants are tested in a large room with a large presentation-size display, they stand, and they are positioned a few meters from the display. The other participants are engaged to act as an audience by attending and sitting around tables in the room during testing. There is no doubt this environment improves external validity. But what about internal validity? Some participants may be distracted or intimidated by the audience. Others might have a tendency to show off, impress, or act out. Such behaviors introduce sources of variation outside the realm of the devices under test, and thereby compromise internal validity. So our effort to improve external validity through environmental considerations may negatively impact internal validity.

A second example pertains to the experimental procedure. Consider an experiment comparing two methods of text entry. In an attempt to improve external validity, participants are instructed to enter whatever text they think of. The text may include punctuation symbols and uppercase and lowercase characters, and participants can edit the text and correct errors as they go. Again, external validity is improved since this is what people normally do when entering text. However, internal validity is compromised because behaviors are introduced that are not directly related to the text entry techniques—behaviors such as pondering (What should

I enter next?) and fiddling with commands (How do I move the cursor back and make a correction? How is overtype mode invoked?). Furthermore, since participants generate the text, errors are difficult to record since there is no "source text" with which to compare the entered text. So here again we see the compromise. The desire to improve external validity through procedural considerations may negatively impact internal validity.

Unfortunately, there is no universal remedy for the tension between internal and external validity. At the very least, one must acknowledge the limitations. Formulating conclusions that are broader than what the results suggest is sure to raise the ire of reviewers. We can strive for the best of both worlds with a simple approach, however. Posing multiple narrow (testable) questions that cover the range of outcomes influencing the broader (untestable) questions will increase both internal and external validity. For example, a technique that is fast, accurate, easy to learn, easy to remember, and considered comfortable and enjoyable by users is generally better. Usually there is a positive correlation between the testable and untestable questions; i.e., participants generally find one UI better than another if it is faster and more accurate, takes fewer steps, is more enjoyable, is more comfortable, and so on.

Before moving on, it is worth mentioning *ecological validity*, a term closely related to external validity. The main distinction is in how the terms are used. Ecological validity refers to the methodology (using materials, tasks, and situations typical of the real world), whereas external validity refers to the outcome (obtaining results that generalize to a broad range of people and situations).

4.7 Comparative evaluations

Evaluating new ideas for user interfaces or interaction techniques is central to research in human-computer interaction. However, evaluations in HCI sometimes focus on a single idea or interface. The idea is conceived, designed, implemented, and evaluated—but not compared. The research component of such an evaluation is questionable. Or, to the extent the exercise is labeled research, it is more aligned with the second definition of research noted earlier: "collecting information about a particular subject."

From a research perspective, our third definition is more appealing, since it includes the ideas of experimentation, discovery, and developing theories of interaction. Certainly, more meaningful and insightful results are obtained if a *comparative evaluation* is performed. In other words, a new user interface or interaction technique is designed and implemented and then compared with one or more alternative designs to determine which is faster, more accurate, less confusing, more preferred by users, etc. The alternatives may be variations in the new design, an established design (a baseline condition), or some combination of the two. In fact, the testable research questions above are crafted as comparisons (e.g., "Is Method A faster than Method B for …?"), and for good reason. A controlled experiment must include at least one independent variable and the independent variable must have at

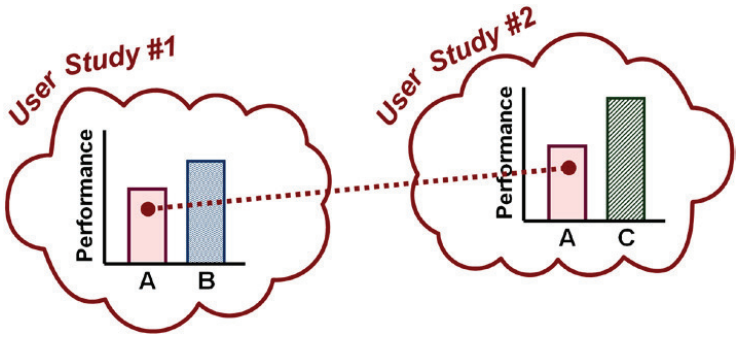

FIGURE 4.10

Including a baseline condition serves as a check on the methodology and facilitates the comparison of results between user studies.

least two levels or test conditions. Comparison, then, is inherent in research following the experimental method discussed earlier. The design of HCI experiments is elaborated further in Chapter 5.

The idea of including an established design as a baseline condition is particularly appealing. There are two benefits. First, the baseline condition serves as a check on the methodology. Baseline conditions are well traveled in the research literature, so results in a new experiment are expected to align with previous results. Second, the baseline condition allows results to be compared with other studies. The general idea is shown in Figure 4.10. The results from two hypothetical user studies are shown. Both user studies are comparative evaluations and both include condition A as a baseline. Provided the methodology was more or less the same, the performance results in the two studies should be the same or similar for the baseline condition. This serves not only as a check on the methodology but also facilitates comparisons between the two user studies. A quick look at the charts suggests that condition C out-performs condition B. This is an interesting observation because condition C was evaluated in one study, condition B in another.

Consider the idea cited earlier of comparing two remote pointing devices for presentation systems. Such a study would benefit by including a conventional mouse as a baseline condition.[8] If the results for the mouse are consistent with those found in other studies, then the methodology was probably okay, and the results for the remote pointing devices are likely valid. Furthermore, conclusions can often be expressed in terms of the known baseline condition, for example, "Device A was found to be about 8 percent slower than a conventional mouse."

The value in conducting a comparative study was studied in research by Tohidi et al. (2006), who tested the hypothesis that a comparative evaluation yields more insight than a one-of evaluation. In their study, participants were assigned to groups and were asked to manually perform simple tasks with climate control interfaces

[8]The example cited earlier on remote pointing devices included a conventional mouse as a baseline condition (MacKenzie and Jusoh, 2001).

(i.e., thermostats). There were three different interfaces tested. Some of the participants interacted with just one interface, while others did the same tasks with all three interfaces. The participants interacting with all three interfaces consistently found more problems and were more critical of the interfaces. They were also less prone to inflate their subjective ratings. While this experiment was fully qualitative—human performance was not measured or quantified—the message is the same: a comparative evaluation yields more valuable and insightful results than a single-interface evaluation.

4.8 Relationships: circumstantial and causal

I noted above that looking for and explaining interesting relationships is part of what we do in HCI research. Often a controlled experiment is designed and conducted specifically for this purpose, and if done properly a particular type of conclusion is possible. We can often say that the condition manipulated in the experiment *caused* the changes in the human responses that were observed and measured. This is a *cause-and-effect relationship*, or simply a *causal relationship*.

In HCI, the variable manipulated is often a nominal-scale attribute of an interface, such as device, entry method, feedback modality, selection technique, menu depth, button layout, and so on. The variable measured is typically a ratio-scale human behavior, such as task completion time, error rate, or the number of button clicks, scrolling events, gaze shifts, etc.

Finding a causal relationship in an HCI experiment yields a powerful conclusion. If the human response measured is vital in HCI, such as the time it takes to do a common task, then knowing that a condition tested in the experiment reduces this time is a valuable outcome. If the condition is an implementation of a novel idea and it was compared with current practice, there may indeed be reason to celebrate. Not only has a causal relationship been found, but the new idea improves on existing practice. This is the sort of outcome that adds valuable knowledge to the discipline; it moves the state of the art forward.[9] This is what HCI research is all about!

Finding a relationship does not necessarily mean a causal relationship exists. Many relationships are *circumstantial*. They exist, and they can be observed, measured, and quantified. But they are not causal, and any attempt to express the relationship as such is wrong. The classic example is the relationship between smoking and cancer. Suppose a research study tracks the habits and health of a large number of people over many years. This is an example of the correlational method of research mentioned earlier. In the end, a relationship is found between smoking and cancer: cancer is more prevalent in the people who smoked. Is it correct to conclude from the study that smoking *causes* cancer? No. The relationship observed is

[9] Reporting a non-significant outcome is also important, particularly if there is reason to believe a test condition might improve an interface or interaction technique. Reporting a non-significant outcome means that, at the very least, other researchers needn't pursue the idea further.

circumstantial, not causal. Consider this: when the data are examined more closely, it is discovered that the tendency to develop cancer is also related to other variables in the data set. It seems the people who developed cancer also tended to drink more alcohol, eat more fatty foods, sleep less, listen to rock music, and so on. Perhaps it was the increased consumption of alcohol that caused the cancer, or the consumption of fatty foods, or something else. The relationship is circumstantial, not causal. This is not to say that *circumstantial relationships* are not useful. Looking for and finding a circumstantial relationship is often the first step in further research, in part because it is relatively easy to collect data and look for circumstantial relationships.

Causal relationships emerge from controlled experiments. Looking for a causal relationship requires a study where, among other things, participants are selected randomly from a population and are randomly assigned to test conditions. A random assignment ensures that each group of participants is the same or similar in all respects except for the conditions under which each group is tested. Thus, the differences that emerge are more likely due to (*caused by*) the test conditions than to environmental or other circumstances. Sometimes participants are balanced into groups where the participants in each group are screened so that the groups are equal in terms of other relevant attributes. For example, an experiment testing two input controllers for games could randomly assign participants to groups or balance the groups to ensure the range of gaming experience is approximately equal.

Here is an HCI example similar to the smoking versus cancer example: A researcher is interested in comparing multi-tap and predictive input (*T9*) for text entry on a mobile phone. The researcher ventures into the world and approaches mobile phone users, asking for five minutes of their time. Many agree. They answer a few questions about experience and usage habits, including their preferred method of entering text messages. Fifteen multi-tap users and 15 *T9* users are found. The users are asked to enter a prescribed phrase of text while they are timed. Back in the lab, the data are analyzed. Evidently, the *T9* users were faster, entering at a rate of 18 words per minute, compared to 12 words per minute for the multi-tap users. That's 50 percent faster for the *T9* users! What is the conclusion? There is a relationship between method of entry and text entry speed; however, the relationship is circumstantial, not causal. It is reasonable to report what was done and what was found, but it is wrong to venture beyond what the methodology gives. Concluding from this simple study that *T9* is faster than multi-tap would be wrong. Upon inspecting the data more closely, it is discovered that the *T9* users tended to be more tech-savvy: they reported considerably more experience using mobile phones, and also reported sending considerably more text messages per day than the multi-tap users who, by and large, said they didn't like sending text messages and did so very infrequently.[10] So the difference observed may be due to prior experience and usage habits, rather than to inherent differences in the text entry methods. If there is a genuine interest in determining if one text entry method

[10] Although it is more difficult to determine, perhaps technically savvy users were more willing to participate in the study. Perhaps the users who declined to participate were predominantly multi-tap users.

is faster than another, a controlled experiment is required. This is the topic of the next chapter.

One final point deserves mention. Cause and effect conclusions are not possible in certain types of controlled experiments. If the variable manipulated is a *naturally occurring attribute* of participants, then cause and effect conclusions are unreliable. Examples of naturally occurring attributes include gender (female, male), personality (extrovert, introvert), handedness (left, right), first language (e.g., English, French, Spanish), political viewpoint (left, right), and so on. These attributes are legitimate independent variables, but they cannot be manipulated, which is to say, they cannot be assigned to participants. In such cases, a cause and effect conclusion is not valid because is not possible to avoid confounding variables (defined in Chapter 5). Being a male, being an extrovert, being left-handed, and so on always brings forth other attributes that systematically vary across levels of the independent variable. Cause and effect conclusions are unreliable in these cases because it is not possible to know whether the experimental effect was due to the independent variable or to the confounding variable.

4.9 Research topics

Most HCI research is not about designing products. It's not even about designing applications for products. In fact, it's not even about design or products. Research in HCI, like in most fields, tends to nip away at the edges. The march forward tends to be incremental. The truth is, most new research ideas tend to build on existing ideas and do so in modest ways. A small improvement to this, a little change to that. When big changes do arise, they usually involve bringing to market, through engineering and design, ideas that already exist in the research literature. Examples are the finger flick and two-finger gestures used on touchscreen phones. Most users likely encountered these for the first time with the Apple iPhone. The gestures seem like bold new advances in interaction, but, of course, they are not. The flick gesture dates at least to the 1960s. Flicks are clearly seen in use with a light pen in the videos of Sutherland's Sketchpad, viewable on YouTube. They are used to terminate a drawing command. Two-finger gestures date at least to the 1970s. Figure 4.11 shows Herot and Weinzapfel's (1978) two-finger gesture used to rotate a virtual knob on a touch-sensitive display. As reported, the knob can be rotated to within 5 degrees of a target position. So what might seem like a bold new advance is often a matter of good engineering and design, using ideas that already exist.

Finding a *research topic* is often the most challenging step for graduate students in HCI (and other fields). The expression "ABD" for "all but dissertation" is a sad reminder of this predicament. Graduate students sometimes find themselves in a position of having finished all degree requirements (e.g., coursework, a teaching practicum) without nailing down the big topic for dissertation research. Students might be surprised to learn that seasoned researchers in universities and industry also struggle for that next big idea. Akin to writer's block, the harder one tries, the

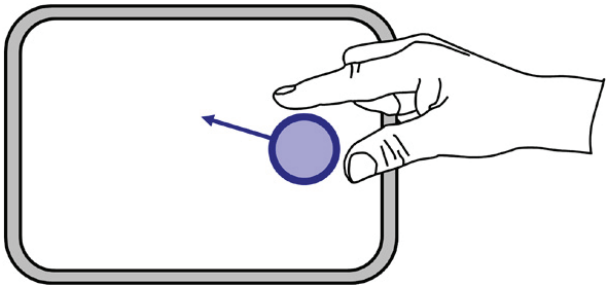

FIGURE 4.11

A two-finger gesture on a touch-sensitive display is used to rotate a virtual knob.

(Adapted from Herot and Weinzapfel, 1978)

less likely is the idea to appear. I will present four tips to overcome "researcher's block" later in this section. First, I present a few observations on ideas and how and where they arise.

4.9.1 Ideas

In the halcyon days after World War II, there was an American television show, a situation comedy, or sitcom, called *The Many Loves of Dobie Gillis* (1959–1963). Much like *Seinfeld* many years later, the show was largely about, well, nothing. Dobie's leisurely life mostly focused on getting rich or on endearing a beautiful woman to his heart. Each episode began with an idea, a scheme. The opening scene often placed Dobie on a park bench beside *The Thinker*, the bronze and marble statue by French sculptor Auguste Rodin (1840–1917). (See Figure 4.12.) After some pensive moments by the statue, Dobie's idea, his scheme, would come to him. It would be nice if research ideas in HCI were similarly available and with such assurance as were Dobie's ideas. That they are not is no cause for concern, however. Dobie's plans usually failed miserably, so we might question his approach to formulating his plans. Is it possible that *The Thinker*, in his pose, is more likely to inspire writer's block than the idea so desperately sought? The answer may be *yes*, but there is little science here. We are dealing with human thought, inspiration, creativity, and a milieu of other human qualities that are poorly understood, at best.

If working hard to find a good idea doesn't work, perhaps a better approach is to relax and just get on with one's day. This seems to have worked for the ancient Greek scholar Archimedes (287–212 BC) who is said to have effortlessly come upon a brilliant idea as a solution to a problem. As a scientist, Archimedes was called upon to determine if King Hiero's crown was pure gold or if it was compromised with a lesser alloy. One solution was to melt the crown, separating the constituent parts. This would destroy the crown–not a good idea. Archimedes' idea was simple, and he is said to have discovered it while taking a bath. Yes, taking a bath, rather than sitting for hours in *The Thinker*'s pose. He realized–in an instant–that the volume of water displaced as he entered the bathtub must equal the volume

FIGURE 4.12

Rodin's *The Thinker* often appeared in the opening scenes of the American sitcom *The Many Loves of Dobie Gillis*.

of his body. Immersing the crown in water would similarly yield the crown's volume, and this, combined with the crown's weight, would reveal the crown's density. If the density of the crown equaled the known density of gold, the King's crown was pure gold–problem solved. According to the legend, Archimedes was so elated at his moment of revelation that he jumped from his bath and ran nude about the streets of Syracuse shouting "Eureka!" ("I found it!").

While legends make good stories, we are not likely to be as fortunate as Archimedes in finding a good idea for an HCI research topic. Inspiration is not always the result of single moment of revelation. It is often gradual, with sources unknown or without a conscious and recognizable connection to the problem. Recall Vannevar Bush's memex, described in the opening chapter of this book. Memex was a concept. It was never built, even though Bush described the interaction with memex in considerable detail. We know memex today as hypertext and the World Wide Web. But where and how did Bush get his idea? The starting point is having a problem to solve. The problem of interest to Bush was coping with ever-expanding volumes of information. Scientists like Bush needed a convenient way to access this information. But how? It seems Bush's inspiration for memex came from… Let's pause for a moment, lest we infer Bush was engaged in a structured approach to problem solving. It is not likely that Bush went to work one morning intent on solving the problem of information access. More than likely, the idea came without deliberate effort. It may have come flittingly, in an instant, or gradually, over days, weeks, or months. Who knows? What is known, however, is that the idea did not arise from nothing. Ideas come from the human experience. This is why in HCI we often read about things like "knowledge in the head and knowledge in the world"

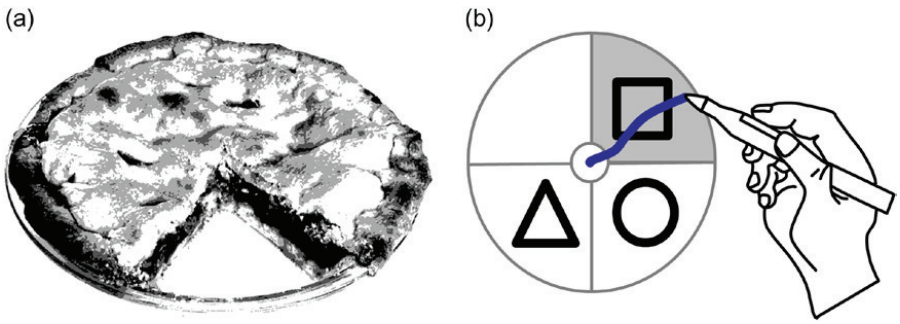

FIGURE 4.13

Pie menus in HCI: (a) The inspiration? (b) HCI example.

(Adapted from G. Kurtenbach, 1993)

(Norman, 1988, ch. 3) or metaphor and analogy (Carroll and Thomas, 1982). The context for inspiration is the human experience. So what was the source of Bush's inspiration for memex? The answer is in Bush's article, and also in Chapter 1.

Are there other examples relevant to HCI? Sure. Twitter co-founder Jack Dorsey is said to have come up with the idea for the popular micro-blogging site while sitting on a children's slide in a park eating Mexican food.[11] What about pie menus in graphical user interfaces? Pie menus, as an alternative to linear menus, were first proposed by Don Hopkins at the University of Maryland in 1988 (cited in Callahan et al., 1988). We might wonder about the source of Hopkins' inspiration (see Figure 4.13).

See also student exercises 4-2 and 4-3 at the end of this chapter.

4.9.2 Finding a topic

It is no small feat to find an interesting research topic. In the following paragraphs, four tips are offered on finding a topic suitable for research. As with the earlier discussion on the cost and frequency of errors (see Figure 3-46), there is little science to offer here. The ideas follow from personal experience and from working with students and other researchers in HCI.

4.9.3 Tip #1: Think small!

At a conference recently, I had an interesting conversation with a student. He was a graduate student in HCI. "Have you found a topic for your research," I asked. "Not really," he said. He had a topic, but only in a broad sense. Seems his supervisor had funding for a large research project related to aviation. The topic, in a general sense, was to develop an improved user interface for an air traffic control system. He was stuck. Where to begin? Did I have any ideas for him? Well, actually, no I didn't. Who wouldn't be stuck? The task of developing a UI for an air traffic control system is huge. Furthermore, the project mostly involves engineering and

[11] *New York Times* Oct 30, 2010, p BU1.

design. Where is the research in designing an improved system of any sort? What are the research questions? What are the experimental variables? Unfortunately, graduate students are often saddled with similar big problems because a supervisor's funding source requires it. The rest of our discussion focused on narrowing the problem—in a big way. Not to some definable sub-system, but to a small aspect of the interface or interaction. The smaller, the better.

The point above is to think small. On finding that big idea, the advice is… forget it. Once you shed that innate desire to find something really significant and important, it's amazing what will follow. If you have a small idea, something that seems a little useful, it's probably worth pursuing as a research project. Pursue it and the next thing you know, three or four related interaction improvements come to mind. Soon enough, there's a dissertation topic in the works. So don't hesitate to think small.

4.9.4 Tip #2: Replicate!

An effective way to get started on research is to replicate an existing experiment from the HCI literature. This seems odd. Where is the research in simply replicating what has already been done? Of course, there is none. But there is a trick. Having taught HCI courses many times over many years, I know that students frequently get stuck finding a topic for the course's research project. Students frequently approach me for suggestions. If I have an idea that seems relevant to the student's interests, I'll suggest it. Quite often (usually!) I don't have any particular idea. If nothing comes to mind, I take another approach. The student is advised just to study the HCI literature—research papers from the CHI proceedings, for example—and find some experimental research on a topic of interest. Then just replicate the experiment. Is that okay, I am asked. Sure, no problem.

The trick is in the path to replicating. Replicating a research experiment requires a lot of work. The process of studying a research paper and precisely determining what was done, then implementing it, testing it, debugging it, doing an experiment around it, and so on will empower the student—the researcher—with a deep understanding of the issues, much deeper than simply reading the paper. This moves the line forward. The stage is set. Quite often, a new idea, a new twist, emerges. But it is important not to *require* something new. The pressure in that may backfire. Something new may emerge, but this might not happen until late in the process, or after the experiment is finished. So it is important to avoid a requirement for novelty. This is difficult, because it is germane to the human condition to strive for something new. Self-doubt may bring the process to a standstill. So keep the expectations low. A small tweak here, a little change there. Good enough. No pressure. Just replicate. You may be surprised with the outcome.

4.9.5 Tip #3: Know the literature!

It might seem obvious, but the process of reviewing research papers on a topic of interest is an excellent way to develop ideas for research projects. The starting

point is identifying the topic in a general sense. If one finds gaming of interest, then gaming is the topic. If one finds social networking of interest, then that's the topic. From there the task is to search out and aggressively study and analyze all published research on the topic. If there are too many publications, then narrow the topic. What, in particular, is the interest in gaming or social networking? Continue the search. Use Google Scholar, the ACM Digital Library, or whatever resource is conveniently available. Download all the papers, store them, organize them, study them, make notes, then open a spreadsheet file and start tabulating features from the papers. In the rows, identify the papers. In the columns, tabulate aspects of the interface or interaction technique, conditions tested, results obtained, and so on. Organize the table in whatever manner seems reasonable.

The process is chaotic at first. Where to begin? What are the issues? The task is daunting, at the very least, because of the divergence in reporting methods. But that's the point. The gain is in the process—bringing shape and structure to the chaos. The table will grow as more papers are found and analyzed. There are examples of such tables in published papers, albeit in a condensed summary form. Figure 4.14 shows an example from a research paper on text entry using small keyboards. The table amounts to a mini literature review. Although the table is neat and tidy, don't be fooled. It emerged from a difficult and chaotic process of reviewing a collection of papers and finding common and relevant issues. The collection of notes in the right-hand column is evidence of the difficulty. This column is like a disclaimer, pointing out issues that complicate comparisons of the data in the other columns.

Are there research topics lurking within Figure 4.14? Probably. But the point is the process, not the product. Building such a table shapes the research area into relevant categories of inquiry. Similar tables are found in other research papers (e.g., Figure 11 and Figure 12 in MacKenzie, 1992; Table 3 and Table 4 in Soukoreff and MacKenzie, 2004). See also student exercise 4-4 at the end of this chapter.

4.9.6 Tip #4: Think inside the box!

The common expression "think outside the box" is a challenge to all. The idea is to dispense with accepted beliefs and assumptions (in the box) and to think in a new way that assumes nothing and challenges everything. However, there is a problem with the challenge. Contemporary, tech-savvy people, clever as they are, often believe they in fact do think outside the box, and that it is everyone else who is confined to life in the box. With this view, the challenge is lost before starting. If there is anything useful in tip #4, it begins with an unsavory precept: You are inside the box! All is not lost, however. Thinking inside the box, then, is thinking about and challenging one's own experiences—the experiences inside the box. The idea is simple. Just get on with your day, but at every juncture, every interaction, think and question. What happened? Why did it happen? Is there an alternative? Play the

Study (1st author)	Number of Keys[a]	Direct/ Indirect	Scanning	Number of Participants	Speed[b] (wpm)	Notes
Bellman [2]	5	Indirect	No	11	11	4 cursors keys + SELECT key. Error rates not reported. No error correction method.
Dunlop [4]	4	Direct	No	12	8.90	4 letter keys + SPACE key. Error rates reported as "very low."
Dunlop [5]	4	Direct	No	20	12	4 letter keys + 1 key for SPACE/NEXT. Error rates not reported. No error correction method.
Tanaka-Ishii [25	3	Direct	No	8	12+	4 letters keys + 4 keys for editing, and selecting. 5 hours training. Error rates not reported. Errors corrected using CLEAR key.
Gong [7]	3	Direct	No	32	8.01	3 letter keys + two additional keys. Error rate = 2.1%. Errors corrected using DELETE key.
MacKenzie [16]	3	Indirect	No	10	9.61	2 cursor keys + SELECT key. Error rate = 2.2%. No error correction method.
Baljko [1]	2	Indirect	Yes	12	3.08	1 SELECT key + BACKSPACE key. 43 virtual keys. RC scanning. Same phrase entered 4 times. Error rate = 18.5%. Scanning interval = 750 ms.
Simpson [24]	1	Indirect	Yes	4	4.48	1 SELECT key. 26 virtual keys. RC scanning. Excluded trials with selection errors or missed selections. No error correction. Scanning interval = 525 ms at end of study.
Koester [10]	1	Indirect	Yes	3	7.2	1 SELECT key. 33 virtual keys. RC scanning with word prediction. Dictionary size not given. Virtual BACKSPACE key. 10 blocks of trials. Error rates not reported. Included trials with selection errors or missed selections. Fastest participant: 8.4 wpm.

[a] For "direct" entry, the value is the number of letter keys. For "indirect" entry, the value is the total number of keys.
[b] The entry speed cited is the highest of the values reported in each source, taken from the last block if multiple blocks.

FIGURE 4.14

Table showing papers (rows) and relevant conditions or results (columns) from research papers on text entry using small keyboards.

(From MacKenzie, 2009b, Table 1; consult for full details on studies cited)

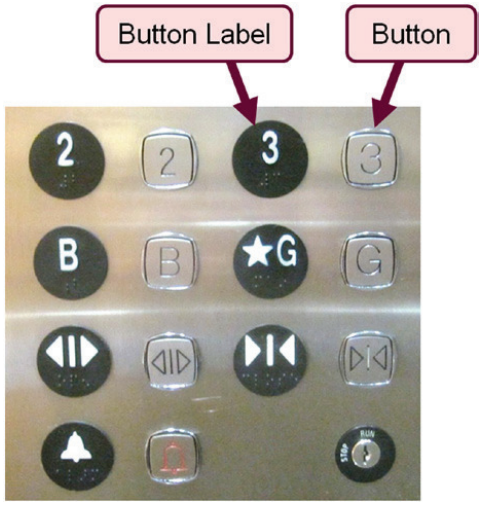

FIGURE 4.15

Elevator control panel. The button label is more prominent than the button.

role of both a participant (this is unavoidable) and an observer. Observe others, of course, but more importantly observe yourself. You are in the box, but have a look, study, and reconsider.

Here's an example, which on the surface seems trivial (but see tip #1). Recently, while at work at York University, I was walking to my class on Human-Computer Interaction. Being a bit late, I was in a hurry. The class was in a nearby building on the third floor and I was carrying some equipment. I entered the elevator and pushed the button—the wrong button. Apparently, for each floor the control panel has both a button label and a button. (See Figure 4.15.) I pushed the button label instead of the button. A second later I pushed the button, and my journey continued. End of story.

Of course, there is more. Why did I push the wrong button? Yes, I was in a hurry, but that's not the full reason. With a white number on a black background, the floor is identified more prominently by the button label than by the button. And the button label is round, like a button. On the button, the number is recessed in the metal and is barely visible. The error was minor, only a *slip* (right intention, wrong action; see Norman, 1988, ch. 5). Is there a research topic in this? Perhaps. Perhaps not. But experiencing, observing, and thinking about one's interactions with technology can generate ideas and promote a humbling yet questioning frame of thinking—thinking that moves forward into research topics. The truth is, I have numerous moments like this every day (and so do you!). Most amount to nothing, but the small foibles in interacting with technology are intriguing and worth thinking about.

In this chapter, we have examined the scientific foundations for research in human-computer interaction. With this, the next challenge is in designing

and conducting experiments using human participants (users) to evaluate new ideas for user interfaces and interaction techniques. We explore these topics in Chapter 5.

STUDENT EXERCISES

4-1. Examine some published papers in HCI and find examples where results were reported as a raw count (e.g., number of errors) rather than as a count *per something* (e.g., percent errors). Find three examples and write a brief report (or prepare a brief presentation) detailing how the results were reported and the weakness or limitation in the method. Propose a better way to report the same results. Use charts or graphs where appropriate.

4-2. What, in Vannevar Bush's "human experience," formed the inspiration for memex? (If needed, review Bush's essay "As We May Think," or see the discussion in Chapter 1.) What are the similarities between his inspiration and memex?

4-3. A *fisheye lens* or *fisheye view* is a tool or concept in HCI whereby high-value information is presented in greater detail than low-value information. Furnas first introduced the idea in 1986 (Furnas, 1986). Although the motivation was to improve the visualization of large data sets, such as programs or databases, Furnas' idea came from something altogether different. What was Furnas' inspiration for fisheye views? Write a brief report describing the analogy offered by Furnas. Include in your report three examples of fisheye lenses, as described and implemented in subsequent research, noting in particular the background motivation.

4-4. Here are some research themes: 3D gaming, mobile phone use while driving, privacy in social networking, location-aware user interfaces, tactile feedback in pointing and selecting, multi-touch tabletop interaction. Choose one of these topics (or another) and build a table similar to that in Figure 4.14. Narrow the topic, if necessary (e.g., mobile phone *texting* while driving), and find at least five relevant research papers to include in the table. Organize the table identifying the papers in the rows and methods, relevant themes, and findings in the columns. Write a brief report about the table. Include citations and references to the selected papers.

4-5. In Chapter 3, we used a 2D plot to illustrate the trade-off between the frequency of errors (x-axis) and the cost of errors (y-axis) (see Figure 3.46). The plot was just a sketch, since the analysis was informal. In this chapter, we discussed another trade-off, that between form and function. The