CHAPTER

# Designing HCI Experiments 5

Learning how to design and conduct an experiment with human participants is a skill required of all researchers in human-computer interaction. In this chapter I describe the core details of designing and conducting HCI experiments.

One way to think about experiment design is through a signal and noise metaphor. In the metaphor, we divide our observations and measurements into two components: signal and noise. (See Figure 5.1.) The source shows a time series. A slight upward trend is apparent; however, the variability or noise in the source makes this difficult to detect. If we separate the source into components for signal and noise, the trend in the signal is clear.

In HCI experiments, the signal is related to a variable of interest, such as input device, feedback mode, or an interaction technique under investigation. The noise is everything else—the random influences. These include environmental circumstances such as temperature, lighting, background noise, a wobbly chair, or glare on the computer screen. The people or participants in the experiment are also a source of noise or variability. Some participants may be having a good day, while others are having a bad day. Some people are predisposed to behave in a certain manner; others behave differently. The process of designing an experiment is one of enhancing the signal while reducing the noise. This is done by carefully considering the setup of the experiment in terms of the variables manipulated and measured, the variables controlled, the procedures, the tasks, and so on. Collectively, these properties of an experiment establish the methodology for the research.

## 5.1 What methodology?

The term *method* or *methodology* refers to the way an experiment is designed and carried out. This involves deciding on the people (participants), the hardware and software (materials or apparatus), the tasks, the order of tasks, the procedure for briefing and preparing the participants, the variables, the data collected and analyzed, and so on. Having a sound methodology is critical. On this point, Allen Newell did not hesitate: "Science is method. Everything else is commentary."[1]

[1]This quote from Allen Newell was cited and elaborated on by Stuart Card in an invited talk at the ACM's SIGCHI conference in Austin, Texas (May 10, 2012).

**Human-Computer Interaction.**

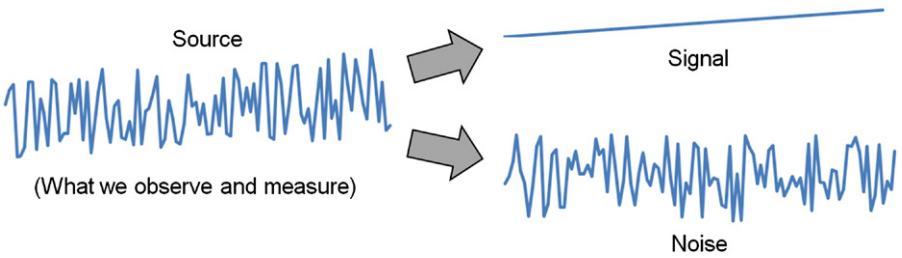



**FIGURE 5.1**

Signal-to-noise conceptualization of experiment design.

These are strong words. Why did Newell apply such forceful yet narrow language to a topic as broad as science? The reason is that Newell, and others, understand that methodology is the bedrock of science. If the methodology is weak or flawed, there is no science forthcoming. What remains is little else than commentary.

In the preceding chapter, I advocated the use of a standardized methodology to strengthen experimental research. The flip side is that an ad hoc, or made-up, methodology weakens research. There is little sense in contriving a methodology simply because it seems like a good way to test or demonstrate an idea. So what is the appropriate methodology for research in human-computer interaction? The discussions that follow pertain only to experimental research and in particular to *factorial experiments*, where participants are exposed to levels of factors (test conditions) while their behavior (human performance) is observed and measured. By and large, the methodology is plucked from one of HCI's parent disciplines: experimental psychology.

Just as the Association for Computing Machinery (ACM) is the dominant organization overseeing computer science and related special interests such as HCI, the American Psychological Association (APA) is the dominant organization overseeing experimental psychology. Their *Publication Manual of the American Psychological Association*, first published in 1929, is a valuable resource for researchers undertaking experimental research involving human participants (APA, 2010). The manual, now in its sixth edition, is used by over 1,000 journals across many disciplines (Belia, Fidler, Williams, and Cumming, 2005). These include HCI journals. The APA guidelines are recommended by journals such as the ACM's *Transactions on Computer-Human Interaction* (*TOCHI*) (ACM, 2012) and Taylor and Francis's *Human-Computer Interaction* (Taylor and Francis, 2012). The APA *Publication Manual* is about more than publishing style; the manual lays out many methodological issues, such as naming and referring to independent and dependent variables, recruiting participants, reporting the results of statistical tests, and so on. Also the important link between research and publication is reflected in the title.

Another resource is psychologist David Martin's *Doing Psychology Experiments*, now in its sixth edition (D. W. Martin, 2004). Martin's approach is

refreshing and entertaining, more cookbook-like than academic. All the core details are there, with examples that teach and amuse.

The proceedings of the ACM SIGCHI's annual conference (CHI) are also an excellent resource. CHI papers are easily viewed and downloaded from the ACM Digital Library. Of course, many research papers in the CHI proceedings do not present experimental research. And that's fine. HCI is multidisciplinary. The research methods brought to bear on human interaction with technology are equally diverse. However, of those papers that do present a *user study*—an experiment with human participants—there are, unfortunately, many where the methodology is ad hoc. The additional burden of weaving one's way through an unfamiliar methodology while simultaneously trying to understand a new and potentially interesting idea makes studying these papers difficult. But there are many CHI papers that stick to the standard methodology for experiments with human participants. It is relatively easy to spot examples. If the paper has a section called Method or Methodology and the first sub-section within is called Participants, there is a good chance the paper and the research it describes follow the standards for experimental research as laid out in this chapter. Examples from the CHI proceedings include the following: Aula, Khan, and Guan, 2010; Chin and Fu, 2010; Chin, Fu, and Kannampallil, 2009; Duggan and Payne, 2008; Gajos, Wobbrock, and Weld, 2008; Kammerer, Nairn, Pirolli, and Chi, 2009; Majaranta, Ahola, and Špakov, 2009; Räihä and Špakov, 2009; Sjölie et al., 2010; Sun, Zhang, Wiedenbeck, and Chintakovid, 2006; Tohidi et al., 2006; Wobbrock et al., 2009.

## 5.2 Ethics approval

One crucial step that precedes the design of every HCI experiment is *ethics approval*. Since HCI research involves humans, "researchers must respect the safety, welfare, and dignity of human participants in their research and treat them equally and fairly."[2] The approval process is governed by the institution or funding agency overseeing the research. At this author's institution, research projects must be approved by the Human Participant Review Committee (HRPC). Other committee names commonly used are the Institutional Review Board (IRB), Ethics Review Committee (ERC), and so on.

Typically, the review committee serves to ensure a number of ethical guidelines are acknowledged and adhered to. These include the right of the participant to be informed of the following:

- The nature of the research (hypotheses, goals and objectives, etc.)
- The research methodology (e.g., medical procedures, questionnaires, participant observation, etc.)
- Any risks or benefits

[2] www.yorku.ca/research/support/ethics/humans.html.

- The right not to participate, not to answer any questions, and/or to terminate participation at any time without prejudice (e.g., without academic penalty, withdrawal of remuneration, etc.)
- The right to anonymity and confidentiality

Details will vary according to local guidelines. Special attention is usually given for vulnerable participants, such as pregnant women, children, or the elderly. The basis for approving the research, where human participants are involved, is in achieving a balance between the risks to participants and the benefits to society.

## 5.3 Experiment design

Experiment design is the process of bringing together all the pieces necessary to test hypotheses on a user interface or interaction technique. It involves deciding on and defining which variables to use, what tasks and procedure to use, how many participants to use and how to solicit them, and so on.

One of the most difficult steps in designing an HCI experiment is just getting started. Ideas about a novel interface or interaction technique take shape well before thoughts of doing an experiment. There may even be an existing prototype that implements a research idea. Perhaps there is no prototype—yet. Regardless, there is an idea about an interface or interaction technique, and it seems new and interesting. Doing an experiment to test the idea seems like a good idea, but it is difficult transitioning from the creative and exciting work of developing a novel idea to the somewhat mechanical and mundane work of doing an experiment. Here is a question that will focus the mind more than any other in getting off and running with an HCI experiment: *what are the experimental variables?*

This seems like an odd place to begin. After all, experimental variables are in a distant world from the creative effort invested thus far. Well, not really. Thinking about experimental variables is an excellent exercise. Here's why. The process forces us to transition from well-intentioned, broad yet untestable questions (e.g., *Is my idea any good?*) to narrower yet testable questions (e.g., *Can a task be performed more quickly with my new interface than with an existing interface?*). If necessary, review the discussion in the preceding chapter on research questions and internal and external validity.

Thinking about experimental variables forces us to craft narrow and testable questions. The two most important experimental variables are *independent variables* and *dependent variables*. In fact, these variables are found within the example question in the preceding paragraph. Expressions like "more quickly" or "fewer steps" capture the essence of dependent variables: human behaviors that are measured. The expression "with my new interface than with an existing interface" captures the essence of an independent variable: an interface that is compared with an alternative interface. In fact, a testable research question inherently expresses the relationship between an independent variable and a dependent variable. Let's examine these two variables in more detail.

## 5.4 Independent variables

An *independent variable* is a circumstance or characteristic that is manipulated or systematically controlled to a change in a human response while the user is interacting with a computer. An independent variable is also called a *factor*. Experiments designed with independent variables are often called *factorial experiments*. The variable is manipulated across multiple (at least two) levels of the circumstance or characteristic. The variable is "independent" because it is independent of participant behavior, which means there is nothing a participant can do to influence an independent variable.

The variable manipulated is typically a nominal-scale attribute, often related to a property of an interface. Review any HCI research paper that presents a factorial experiment and examples of independent variables are easily found. They are anything that might affect users' proficiency in using a computer system. Examples include device (with levels mouse, trackball, and stylus) (MacKenzie, Sellen, and Buxton, 1991), feedback modality (with levels auditory, visual, and tactile) (Akamatsu et al., 1995), display size (with levels large and small) (Dillon, Richardson, and Mcknight, 1990), display type (with levels CRT and LCD) (MacKenzie and Riddersma, 1994), cross-display technique (with levels stitching, mouse ether, and ether + halo) (Nacenta, Regan, Mandry, and Gutwin, 2008), transfer function (with levels constant gain and pointer acceleration) (Casiez, Vogel, Pan, and Chaillou, 2007), tree visualization (with levels traditional, list, and multi-column) (Song, Kim, Lee, and Seo, 2010), and navigation technique (with levels standard pan and zoom versus PolyZoom) (Javed, Ghani, and Elmqvist, 2012). These variables are easy to manipulate because they are attributes of the apparatus (i.e., the computer or software). The idea of "manipulating" simply refers to systematically giving one interface, then another, to participants as part of the experimental procedure.

However, an independent variable can be many things besides an attribute of a computer system. It can be characteristics of humans, such as age (Chin and Fu, 2010; Chin et al., 2009), gender (male, female) (Sun et al., 2006; Zanbaka, Goolkasian, and Hodges, 2006), handedness (left handed, right handed) (Kabbash et al., 1993; Peters and Ivanoff, 1999), expertise in assessing web pages (expert, novice) (Brajnik, Yesilada, and Harper, 2011), body position (standing, sitting, walking), preferred operating system (Windows, Mac OS, Linux), first language (e.g., English, French, Chinese), political viewpoint (left, right), religious viewpoint, highest level of education, income, height, weight, hair color, shoe size, and so on. It is not clear that these human characteristics necessarily relate to HCI, but who knows.

Note that human characteristics such as gender or first language are *naturally occurring attributes*. Although such attributes are legitimate independent variables, they cannot be "manipulated" in the same way as an attribute of an interface.

An independent variable can also be an environmental circumstance, such as background noise (quiet, noisy), room lighting (sun, incandescent, fluorescent), vibration level (calm, in a car, in a train), and so on.

Here are two tips to consider. First, when formulating an independent variable, express it both in terms of the circumstance or characteristic itself as well as the

levels of the circumstance or characteristic chosen for testing. (The levels of an independent variable are often called *test conditions*.) So we might have an independent variable called *interaction stance* with levels *sitting*, *standing*, and *walking*. This might seem like an odd point; however, in reading HCI research papers, it is surprising how often an independent variable is not explicitly named. For example, if an experiment seeks to determine whether a certain PDA task is performed better using audio versus tactile feedback, it is important to separately name both the independent variable (e.g., *feedback modality*) and the levels (*audio*, *tactile*).

The second tip is related to the first: Once the name of the independent variable and the names of the levels are decided, stick with these terms consistently throughout a paper. These terms hold special meaning within the experiment and any deviation in form is potentially confusing to the reader. Switching to terms like interaction position (cf. interaction stance), upright (cf. standing), sound (cf. audio), or vibration (cf. tactile) is potentially confusing. Is this a minor, nit-picky point? No. At times, it is a struggle to follow the discussions in a research paper. The fault often lies in the write-up, not in one's ability to follow or understand. The onus is on the researcher writing up the results of his or her work to deliver the rationale, methodology, results, discussion, and conclusions in the clearest way possible. Writing in a straightforward, consistent, and concise voice cannot be overemphasized. Further tips on writing for clarity are elaborated in Chapter 8.

Although it is reasonable to design and conduct an HCI experiment with a single independent variable, experiments often have more than one independent variable. Since considerable work is invested in designing and executing an experiment, there is a tendency to pack in as many independent variables as possible, so that more research questions are posed and, presumably, answered. However, including too many variables may compromise the entire experiment. With every additional independent variable, more effects exist between the variables. Figure 5.2 illustrates.

A design with a single independent variable includes a *main effect*, but no *interaction effects*. A design with two independent variables includes two main effects and one interaction effect, for a total of three effects. The interaction effect is a *two-way interaction*, since it is between two independent variables. For example, an experiment with independent variables Device and Task includes main effects for

| Independent variables | Effects | | | | | Total |
|---|---|---|---|---|---|---|
| | Main | 2-way | 3-way | 4-way | 5-way | |
| 1 | 1 | - | - | - | - | 1 |
| 2 | 2 | 1 | - | - | - | 3 |
| 3 | 3 | 3 | 1 | - | - | 7 |
| 4 | 4 | 6 | 3 | 1 | - | 14 |
| 5 | 5 | 10 | 6 | 3 | 1 | 25 |

**FIGURE 5.2**

The number of effects (main and interaction) increases as the number of independent variables increases.

Device and Task as well as a Device × Task interaction effect. As a reminder, the effect is *on the dependent variable*. The interpretation of interaction effects is discussed in Chapter 6 on Hypothesis Testing.

Once a third independent variable is introduced, the situation worsens: there are seven effects! With four and five independent variables, there are 14 and 25 effects, respectively. Too many variables! It is difficult to find meaningful interpretations for all the effects where there are so many. Furthermore, variability in the human responses is added with each independent variable, so all may be lost if too many variables are included. Interaction effects that are three-way or higher are extremely difficult to interpret and are best avoided. A good design, then, is one that limits the number of independent variables to one or two, three at most.[3]

## 5.5 Dependent variables

A *dependent variable* is a measured human behavior. In HCI the most common dependent variables relate to speed and accuracy, with speed often reported in its reciprocal form, time—task completion time. Accuracy is often reported as the percentage of trials or other actions performed correctly or incorrectly. In the latter case, accuracy is called errors or error rate. The *dependent* in dependent variable refers to the variable being dependent on the human. The measurements *depend on* what the participant does. If the dependent variable is, for example, task completion time, then clearly the measurements are highly dependent on the participant's behavior.

Besides speed and accuracy, a myriad of other dependent variables are used in HCI experiments. Others include preparation time, action time, throughput, gaze shifts, mouse-to-keyboard hand transitions, presses of BACKSPACE, target re-entries, retries, key actions, gaze shifts, wobduls, etc. The possibilities are limitless.

Now, if you are wondering about "wobduls," then you're probably following the discussion. So what is a wobdul? Well, nothing, really. It's just a made-up word. It is mentioned only to highlight something important in dependent variables: Any observable, measurable aspect of human behavior is a potential dependent variable. Provided the behavior has the ability to differentiate performance between two test conditions in a way that might shed light on the strengths or weaknesses of one condition over another, then it is a legitimate dependent variable. So when it comes to dependent variables, it is acceptable to "roll your own." Of course, it is essential to clearly define all dependent variables to ensure the research can be replicated.

An example of a novel dependent variable is "negative facial expressions" defined by Duh et al. (2008) in a comparative evaluation of three mobile phones used for gaming. Participants were videotaped playing games on different mobile

[3] We should add that additional independent variables are sometimes added simply to ensure the procedure covers a representative range of behaviors. For example, a Fitts' law experiment primarily interested in device and task might also include movement distance and target size as independent variables—the latter two included to ensure the task encompasses a typical range of target selection conditions (MacKenzie et al., 1991).

phones. A post-test analysis of the videotape was performed to count negative facial expressions such as frowns, confusion, frustration, and head shakes. The counts were entered in an analysis of variance to determine whether participants had different degrees of difficulty with any of the interfaces.

Another example is "read text events." In pilot testing a system using an eye tracker for text entry (eye typing), it was observed that users frequently shifted their point of gaze from the on-screen keyboard to the typed text to monitor their progress (Majaranta et al., 2006). Furthermore, there was a sense that this behavior was particularly prominent for one of the test conditions. Thus RTE (read text events) was defined and used as a dependent variable. The same research also used "re-focus events" (RFE) as a dependent variable. RFE was defined as the number of times a participant refocuses on a key to select it.

Unless one is investigating mobile phone gaming or eye typing, it is unlikely negative facial expressions, read text events, or refocus events are used as dependent variables. They are mentioned only to emphasize the merit in defining, measuring, and analyzing any human behavior that might expose differences in the interfaces or interaction techniques under investigation.

As with independent variables, it is often helpful to name the variable separately from its units. For example, in a text entry experiment there is likely a dependent variable called *text entry speed* with units "words per minute." Experiments on computer pointing devices often use a Fitts' law paradigm for testing. There is typically a dependent variable named *throughput* with units "bits per second." The most common dependent variable is *task completion time* with units "seconds" or "milliseconds." If the measurement is a simple count of events, there is no unit per se.

When contriving a dependent variable, it is important to consider how the measurements are gathered and the data collected, organized, and stored. The most efficient method is to design the experimental software to gather the measurements based on time stamps, key presses, or other interactions detectable through software events. The data should be organized and stored in a manner that facilitates follow-up analyses. Figure 5.3 shows an example for a text entry experiment. There are two data files. The first contains timestamps and key presses, while the second summarizes entry of a complete phrase, one line per phrase.

The data files in Figure 5.3 were created through the software that implements the user interface or interaction technique. Pilot testing is crucial. Often, pilot testing is considered a rough test of the user interface—with modifications added to get the interaction right. And that's true. But pilot testing is also important to ensure the data collected are correct and available in an appropriate format for follow-on analyses. So pilot test the experiment software and perform preliminary analyses on the data collected. A spreadsheet application is often sufficient for this.

To facilitate follow-up analyses, the data should also include codes to identify the participants and test conditions. Typically, this information is contained in additional columns in the data or in the filenames. For example, the filename for the data in Figure 5.3a is TextInputHuffman-P01-D99-B06-S01.sd1 and identifies the

(a)

```
my bike has a flat tire
my bike has a flat tire
16 3
891 2
1797 3 m
3656 2
4188 1
4672 2 y
5750 3
5938 3 [Space]
6813 3
6984 2
7219 0
8656 3 b
```

(b)

```
min_keystrokes,keystrokes,presented_characters,transcribed_characters, ...
55, 59, 23, 23, 29.45, 0, 9.37, 0.0, 2.5652173913043477, 93.22033898305085
61, 65, 26, 26, 30.28, 0, 10.3, 0.0, 2.5, 93.84615384615384
85, 85, 33, 33, 48.59, 0, 8.15, 0.0, 2.5757575757575757, 100.0
67, 71, 28, 28, 33.92, 0, 9.91, 0.0, 2.5357142857142856, 94.36619718309859
61, 70, 24, 24, 39.44, 0, 7.3, 0.0, 2.9166666666666665, 87.14285714285714
```

**FIGURE 5.3**

Example data files from a text entry experiment: (a) The summary data one (sd1) file contains timestamps and keystroke data. (b) The summary data two (sd2) file contains one line for each phrase of entry.

experiment (TextInputHuffman), the participant (P01), the device (D99), the block (B06) and the session (S01). The suffix is "sd1" for "summary data one." Note that the sd2 file in Figure 5.3b is comma-delimited to facilitate importing and contains a header line identifying the data in each column below.

If the experiment is conducted using a commercial product, it is often impossible to collect data through custom experimental software. Participants are observed externally, rather than through software. In such cases, *data collection* is problematic and requires a creative approach. Methods include manual timing by the experimenter, using a log sheet and pencil to record events, or taking photos or screen snaps of the interaction as entry proceeds. A photo is useful, for example, if results are visible on the display at the end of a trial. Videotaping is another option, but follow-up analyses of video data are time consuming. Companies such as Noldus (www.noldus.com) offer complete systems for videotaping interaction and performing post hoc timeline analyses.

## 5.6 Other variables

Besides independent and dependent variables, there are three other variables: control, random, and confounding. These receive considerably less attention and are rarely mentioned in research papers. Nevertheless, understanding each is important for experimental research.

### 5.6.1 Control variables

There are many circumstances or factors that (a) might influence a dependent variable but (b) are not under investigation. These need to be accommodated in some manner. One way is to control them—to treat them as *control variables*. Examples include room lighting, room temperature, background noise, display size, mouse shape, mouse cursor speed, keyboard angle, chair height, and so on. Mostly, researchers don't think about these conditions. But they exist and they might influence a dependent variable. Controlling them means that they are fixed at a nominal setting during the experiment so they don't interfere. But they might interfere if set at an extreme value. If the background noise level is very high or if the room is too cold, these factors might influence the outcome. Allowing such circumstances to exist at a fixed nominal value is typical in experiment research. The circumstances are treated as control variables.

Sometimes it is desirable to control characteristics of the participants. The type of interface or the objectives of the research might necessitate testing participants with certain attributes, for example, right-handed participants, participants with 20/20 vision, or participants with certain experience. Having lots of control variables reduces the variability in the measured behaviors but yields results that are less generalizable.

### 5.6.2 Random variables

Instead of controlling all circumstances or factors, some might be allowed to vary randomly. Such circumstances are *random variables*. There is a *cost* since more variability is introduced in the measures, but there is a *benefit* since results are more generalizable.

Typically, random variables pertain to characteristics of the participants, including biometrics (e.g., height, weight, hand size, grip strength), social disposition (e.g., conscientious, relaxed, nervous), or even genetics (e.g., gender, IQ). Generally, these characteristics are allowed to vary at random.

Before proceeding, it is worth summarizing the trade-off noted above for control and random variables. The comparison is best presented when juxtaposed with the experimental properties of internal validity and external validity, as discussed in the preceding chapter. Figure 5.4 shows the trade-off.

### 5.6.3 Confounding variables

Any circumstance or condition that changes systematically with an independent variable is a *confounding variable*. Unlike control or random variables, confounding variables are usually problematic in experimental research. Is the effect observed due to the independent variable or to the confounding variable? Researchers must attune to the possible presence of a confounding variable and eliminate it, adjust for it, or consider it in some way. Otherwise, the effects observed may be incorrectly interpreted.

| Variable | Advantage | Disadvantage |
|---|---|---|
| Random | Improves external validity by using a variety of situations and people. | Compromises internal validity by introducing additional variability in the measured behaviours. |
| Control | Improves internal validity since variability due to a controlled circumstance is eliminated | Compromises external validity by limiting responses to specific situations and people. |

**FIGURE 5.4**

Relationship between random and control variables and internal and external validity.

As an example, consider an experiment seeking to determine if there is an effect of camera distance on human performance using an eye tracker for computer control. In the experiment, camera distance—the independent variable—has two levels, near and far. For the near condition, a small camera (A) is mounted on a bracket attached to the user's eye glasses. For the far condition, an expensive eye tracking system is used with the camera (B) positioned above the system's display. Here, camera is a confounding variable since it varies systematically across the levels of the independent variable: camera A for the near condition and camera B for the far condition. If the experiment shows a significant effect of camera distance on human performance, there is the possibility that the effect has nothing to do with camera distance. Perhaps the effect is simply the result of using one camera for the near condition and a different camera for the far condition. The confound is avoided by using the same camera (and same system) in both the near and far conditions. Another possibility is simply to rename the independent variable. The new name could be "setup," with levels "near setup" and "far setup." The new labels acknowledge that the independent variable encompasses multiple facets of the interface, in this case, camera distance, camera, and system. The distinction is important if for no other reason than to ensure the conclusions speak accurately to the different setups, rather than to camera distance alone.

Confounding variables are sometimes found in Fitts' law experiments. Most Fitts' law experiments use a target selection task with movement amplitude ($A$) and target width ($W$) as independent variables. Fitts' original experiment is a typical example. He used a stylus-tapping task with four levels each for movement amplitude ($A = 0.25$, 0.5, 1.0, and 2.0 inches) and target width ($W = 2$, 4, 8, and 16 inches) (Fitts, 1954).[4] Fitts went beyond simple target selection, however. He argued by analogy with information theory and electronic communications that $A$ and $W$ are like signal and noise, respectively, and that each task carries information in *bits*. He proposed an *index of difficulty* (*ID*) as a measure in bits of the information content of a task: $ID = \log_2(2A/W)$. Although the majority of Fitts' law experiments treat $A$ and $W$ as independent variables, sometimes $A$ and $ID$ are treated as independent variables (e.g., Gan and Hoffmann, 1988). Consider the example in Figure 5.5a. There are two independent variables, $A$ with levels 4, 8, 16, and 32 cm, and $ID$ with

[4] See also section 7.7.7, Fitts' law.

(a)

| ID (bits) | Amplitude (pixels) | | | |
|---|---|---|---|---|
| | 16 | 32 | 64 | 128 |
| 1 | * | * | * | * |
| 2 | * | * | * | * |
| 3 | * | * | * | * |
| 4 | * | * | * | * |

(b)

| ID (bits) | Amplitude (pixels) | | | |
|---|---|---|---|---|
| | 16 | 32 | 64 | 128 |
| 1 | **16** | **32** | **64** | **128** |
| 2 | **8** | **16** | **32** | **64** |
| 3 | **4** | **8** | **16** | **32** |
| 4 | **2** | **4** | **8** | **16** |

(c)

| W (pixels) | Amplitude (pixels) | | | |
|---|---|---|---|---|
| | 16 | 32 | 64 | 128 |
| 2 | * | | | |
| 4 | * | * | | |
| 8 | * | * | * | |
| 16 | * | * | * | * |
| 32 | | * | * | * |
| 64 | | | * | * |
| 128 | | | | * |

**FIGURE 5.5**

Fitts' law experiment with *A* and *ID* as independent variables: (a) Yellow cells (*) show the test conditions. (b) Numbers in yellow cells show the target width, revealing a confounding variable. (c) Same design showing test conditions by *A* and *W*.

levels 1, 2, 3, and 4 bits, yielding $4 \times 4 = 16$ test conditions (shaded/yellow cells in the figure). To achieve the necessary *ID* for each *A*, target width must vary. The top-left cell, for example, requires $W = 2A/2^{ID} = (2 \times 16)/2^1 = 16$ cm. The target width for each condition is added in Figure 5.5b. Do you see the confound? As *ID* increases, *W* decreases. Target width (*W*) is a confounding variable. If the experiment reveals a significant effect of *ID*, is the effect due to *ID* or to *W*?[5] To further illustrate, Figure 5.5c shows the same design, but reveals the conditions by movement amplitude (*A*) and target width (*W*).

As another example, consider an experiment with "interaction technique" as an independent variable with three levels, A, B, and C. Assume, further, that there were 12 participants and all were tested on A, then B, then C. Clearly, performance might improve due to practice. Practice, in this case, is a confounding variable because it changes systematically with interaction technique. Participants had just a little practice for A, a bit more for B, still more for C. If performance was best for C, it would be nice to conclude that C is better than A or B. However, perhaps performance was better simply because participants benefited from practice on

[5]It can be argued that a traditional Fitts' law design, using *A* and *W* as independent variables, is similarly flawed because it contains *ID* as a confounding variable. However, this is a weak argument: *A* and *W* are primitive characteristics of the task whereas *ID* is a contrived variable, calculated from *A* and *W*.

A and B prior to testing on C. One way to accommodate practice as a confounding variable is to counterbalance the order of presenting the test conditions to participants (see section 5.11).

Here is another example: Two search engine interfaces are compared, Google versus "new." If all participants have prior experience with Google but no experience with the new interface, then *prior experience* is a confounding variable. This might be unavoidable, as it is difficult to find participants without experience with the Google search engine. As long as the effect of prior experience is noted and acknowledged, then this isn't a problem. Of course, the effect may be due to the confound, not to the test conditions.

A similar confound may occur in text entry experiments where, for example, a new keyboard layout is compared with a Qwerty layout. A fair comparison would require participants having the same level of experience with both layouts. But of course it is difficult to find participants unfamiliar with the Qwerty layout. Thus the Qwerty layout is certain to have an advantage, at least initially. In such cases, it is worth considering a longitudinal design, where the layouts are compared over a prolonged period to see if the new keyboard layout has the potential to overcome Qwerty with practice.

## 5.7 Task and procedure

Let's revisit the definition of an independent variable: "a circumstance or characteristic that is manipulated or systematically controlled to *elicit a change in a human response* while the user is interacting with a computer." Emphasis is added to "elicit a change in a human response." When participants are given a test condition, they are asked to do a task while their performance is measured. Later, they are given a different test condition—another level of the independent variable—and asked to do the task again. Clearly, the choice of task is important.

There are two objectives in designing a good task: *represent* and *discriminate*. A good task is representative of the activities people do with the interface. A task that is similar to actual or expected usage will improve the external validity of the research–the ability to generalize results to other people and other situations. A good task is also one that can discriminate the test conditions. Obviously, there is something in the interaction that differentiates the test conditions, otherwise there is no research to conduct. A good task must attune to the points of differentiation in order to elicit behavioral responses that expose benefits or problems among the test conditions. This should surface as a difference in the measured responses across the test conditions. A difference might occur if the interfaces or interaction techniques are sufficiently distinct in the way the task is performed.

Often, the choice of a task is self-evident. If the research idea is a graphical method for inserting functions in a spreadsheet, a good task is inserting functions into a spreadsheet—using the graphical method versus the traditional typing method. If the research idea is an auditory feedback technique while programming

a GPS device, a good task is programming a destination in a GPS device—aided with auditory feedback versus visual feedback.

Making a task representative of actual usage will improve external validity, but there is a downside. The more representative the task, the more the task is likely to include behaviors not directly related to the interface or interaction method under test. Such behaviors are likely to compromise the ability of the task to discriminate among the test conditions. There is nothing sinister in this. It is simply a reflection of the complex way humans go about their business while using computers. When we enter text, we also think about what to enter. We might pause, think, enter something, think again, change our minds, delete something, enter some more, and so on. This is actual usage. If the research goal is to evaluate a new text entry method, a task that mimics actual usage is fraught with problems. Actual usage includes secondary tasks—lots of them. If the task involves, for example, measuring text entry speed in words per minute, the measurement is seriously compromised if tasks unrelated to the entry method are present.

While using a task that is representative of actual usage may improve external validity, the downside is a decrease in internal validity. Recall that high internal validity means the effects observed (i.e., the differences in means on a dependent variable) are due to the test conditions. The additional sources of variation introduced by secondary tasks reduce the likelihood that the differences observed are actually due to, or caused by, the test conditions. The differences may simply be artifacts of the secondary tasks. Furthermore, the additional variation may bring forth a non-significant statistical result. This is unfortunate if indeed there are inherent differences between the test conditions—differences that should have produced a statistically significant outcome.

The best task is one that is natural yet focuses on the core aspects of the interaction: the points of differentiation between the test conditions. Points of similarly, while true to actual usage, introduce variability. Consider two different text entry techniques being compared in an experimental evaluation. If the techniques include the same method of capitalization, then capitalization does not serve to discriminate the techniques and can be excluded from the experimental task. Including capitalization will improve external validity but will also compromise internal validity due to the added variability.

The tasks considered above are mostly performance-based or skill-based. Sometimes an independent variable necessitates using a knowledge-based task. For example, if the research is comparing two search methods, a reasonable task is to locate an item of information in a database or on the Internet (e.g., “Find the date of birth of Albert Einstein.”). Performance is still measured; however, the participant acquires knowledge of the task goal and, therefore, is precluded from further exposure to the same task. This is a problem if the independent variable is assigned within-subjects (discussed below). When the participant is tested with the other search method, the task must be changed (e.g., “Find the date of birth of William Shakespeare.”). This is tricky, since the new task must be more or less the same (so the search methods can be compared), but also different enough so that the participant does not benefit from exposure to the earlier, similar task.

The experimental procedure includes the task but also the instructions, demonstration, or practice given to the participants. The procedure encompasses everything that the participant did or was exposed to. If a questionnaire was administered before or after testing, it is also part of the experimental procedure and deserves due consideration and explanation in the write-up of the experiment.

## 5.8 Participants

Researchers often assume that their results apply to people who were not tested. Applying results to people other than those who were tested is possible; however, two conditions are required. First, the people actually tested must be members of the same population of people to whom results are assumed to hold. For example, results are unlikely to apply to children if the participants in the experiment were drawn exclusively from the local university campus. Second, a sufficient number of participants must be tested. This requirement has more to do with statistical testing than with the similarly of participants to the population.[6] Within any population, or any sample drawn from a population, variability is present. When performance data are gathered on participants, the variability in the measurements affects the likelihood of obtaining statistically significant results. Increasing the number of participants (large $n$) increases the likelihood of achieving statistically significant results.

In view of the point above, we might ask: How many participants should be used in an experiment? Although the answer might seem peculiar, it goes something like this: use the same number of participants as in similar research (D. W. Martin, 2004, p. 234). Using more participants seems like a good idea, but there is a downside. If there truly is an inherent difference in two conditions, then it is always possible to achieve statistical significance—if enough participants are used. Sometimes the inherent difference is slight, and therein lies the problem. To explain, here is a research question to consider: *Is there a speed difference between left-handed and right-handed users in performing point-select tasks using a mouse?* There may be a slight difference, but it likely would surface only if a very large number of left- and right-handed participants were tested. Use enough participants and statistically significant results will appear. But the difference may be small and of no practical value. Therein lies the problem of using a large number of participants: statistically significant results for a difference of no practical significance.

The converse is also problematic. If not enough participants are used, statistical significance may fail to appear. There might be a substantial experimental effect, but the variance combined with a small sample size (not enough participants) might prevent statistical significance from appearing.

It is possible to compute the power of statistical tests and thereby determine the number of participants required. The analysis may be done *a priori*—before an

[6] Participants drawn from a population are, by definition, similar to the population, since they (collectively) define the population.

experiment is conducted. In practice, *a priori* power analysis is rarely done because it hinges on knowing the variance in a sample before the data are collected.[7] The recommendation, again, is to study published research. If an experiment similar to that contemplated reported statistically significant results with 12 participants, then 12 participants is a good choice.

In HCI, we often hear of researchers doing *usability evaluation* or *usability testing*. These exercises often seek to assess a prototype system with users to determine problems with the interface. Such evaluations are typically not organized as factorial experiments. So the question of how many participants is not relevant in a statistical sense. In usability evaluations, it is known that a small number of participants is sufficient to expose a high percentage of the problems in an interface. There is evidence that about five participants (often usability experts) are sufficient to expose about 80 percent of the usability problems (Lewis, 1994; Nielsen, 1994).

It is worth reflecting on the term *participants*. When referring specifically to the experiment, use the term participants (e.g., "all participants exhibited a high error rate").[8] General comments on the topic or conclusions drawn may use other terms (e.g., "these results suggest that users are less likely to…").

When recruiting participants, it is important to consider how the participants are selected. Are they solicited by word of mouth, through an e-mail list, using a notice posted on a wall, or through some other means? Ideally, participants are drawn at random from a population. In practice, this is rarely done, in part because of the need to obtain participants that are close by and available. More typically, participants are solicited from a convenient pool of individuals (e.g., members in the workplace, children at a school, or students from the local university campus). Strictly speaking, convenience sampling compromises the external validity of the research, since the true population is somewhat narrower than the desired population.

To help identify the population, participants are typically given a brief questionnaire (discussed shortly) at the beginning or end of the experiment to gather demographic data, such as age and gender. Other information relevant to the research is gathered, such as daily computer usage or experience with certain applications, devices, or products.

HCI experiments often require participants with specific skills. Perhaps a filtering process is used to ensure only appropriate participants are used. For example, an experiment investigating a new gaming input device might want a participant pool with specific skills, such a minimum of 15 hours per week playing computer games. Or perhaps participants without gaming experience are desired. Whatever the case, the selection criteria should be clear and should be stated in the write-up of the methodology, in a section labeled "Participants."

---

[7] Computing power in advance of an experiment also requires the researcher to know the size of the experimental effect (the difference in the means on the dependent variable) that is deemed relevant. Usually, the researcher simply wants to know if there is a statistically significant difference without committing in advance to a particular difference being of practical significance.

[8] According to the APA guidelines, the term *subjects* is also acceptable (APA, 2010, p. 73).

Depending on the agency or institution overseeing the research, participants are usually required to sign a consent form prior to testing. The goal is to ensure participants know that their participation is voluntary, that they will incur no physical or psychological harm, that they can withdraw at any time, and that their privacy, anonymity, and confidentiality will be protected.

## 5.9 Questionnaire design

Questionnaires are a part of most HCI experiments. They have two purposes. One is to gather information on demographics (age, gender, etc.) and experience with related technology. Another is to solicit participants' opinions on the devices or interaction tasks with which they are tested.

Questionnaires are the primary instrument for survey research, a form of research seeking to solicit a large number of people for their opinions and behaviors on a subject such as politics, spending habits, or use of technology. Such questionnaires are often lengthy, spanning several pages. Questionnaires administered in HCI experiments are usually more modest, taking just a few minutes to complete.

Questions may be posed in several ways, depending on the nature of the information sought and how it is to be used. Let's look at a few examples. Closed-ended questions are convenient, since they constrain a participant's response to small set of options. The following are examples of close-ended questions:

Do you use a GPS device while driving? ☐ yes ☐ no

Which browser do you use?
☐ Mozilla *Firefox* ☐ Google *Chrome*
☐ Microsoft *IE* ☐ Other ( _____ )

The question above includes an open-ended category, "Other." Of course, the entire question could be open-ended, as shown here:

Which browser do you use? _____

Closed-end questions simplify follow-on analyses, since it is straightforward to tally counts of responses.

It is usually important to know the gender and age of participants, since this helps identify the population. Age can be solicited as an open-ended ratio-scale response, as seen here:

Please indicate your age:_____

Collected in this manner, the mean and standard deviation are easily calculated. Ratio-scale responses are also useful in looking for relationships in data. For example, if the same questionnaire also included a ratio-scale item on the number of text messages sent per day, then it is possible to determine if the responses are correlated (e.g., *Is the number of text messages sent per day related to age?*). Age can also be solicited as an ordinal response, as in this example:

Please indicate your age.

☐ < 20 ☐ 20-29 ☐ 30-39
☐ 40-49 ☐ 50-59 ☐ 60+

In this case, the counts in each category are tabulated. Such data are particularly useful if there is a large number of respondents. However, ordinal data are inherently lower quality than ratio-scale data, since it is not possible to compute the mean or standard deviation.

Questionnaires are also used at the end of an experiment to obtain participants' opinions and feelings about the interfaces or interaction techniques. Items are often formatted using a Likert scale (see Figure 4.7) to facilitate summarizing and analyzing the responses. One example is the NASA-TLX (task load index), which assesses perceived workload on six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart and Staveland, 1988). A questionnaire on frustration may be presented as follows:

**Frustration**: I felt a high level of insecurity, discouragement, irritation, stress, or annoyance.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| Strongly disagree | | | Neutral | | | Strongly agree |

The ISO 9241-9 standard for non-keyboard input devices includes a questionnaire with 12 items to assess the comfort and fatigue experienced by participants (ISO 2000). The items are similar to those in the NASA-TLX but are generally directed to interaction with devices such as mice, joysticks, or eye trackers. The items may be tailored according to the device under test. For example, an evaluation of an eye tracker for computer control might include a questionnaire with the following response choices (see also Zhang and MacKenzie, 2007):

**Eye fatigue**:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| Very high | | | | | | Very low |

Note that the preferred response is 7, whereas the preferred response in the NASA-TLX example is 1. In the event the mean is computed over several response items, it is important that the items are consistently constructed.

## 5.10 Within-subjects and between-subjects

The administering of test conditions (levels of a factor) is either *within-subjects* or *between-subjects*. If each participant is tested on each level, the assignment is within-subjects. Within-subjects is also called *repeated measures*, because the measurements on each test condition are repeated for each participant. If each participant is tested on only one level, the assignment is between-subjects. For a *between-subjects design*, a separate group of participants is used for each test condition. Figure 5.6 provides a simple illustration of the difference between a within-subjects assignment and a between-subjects assignment. The figure assumes a single factor with three levels: A, B, and C. Figure 5.6a shows a within-subjects assignment because each participant is tested on all three levels of the factor (but see section 5.11, Counterbalancing). Figure 5.6b shows a between-subjects assignment, since each participant is tested on only one level of the factor. There are three groups of participants, with two participants in each group.

Clearly, there is a trade-off. For a between-subjects design, each participant is tested on only one level of a factor; therefore, more participants are needed to obtain the same number of observations (Figure 5.6b). For a *within-subjects design*, each participant is tested on all levels of a factor. Fewer participants are needed; however, more testing is required for each participant (Figure 5.6a). Given this trade-off, it is reasonable to ask: is it better to assign a factor within-subjects or between-subjects? Let's examine the possibilities.

Sometimes a factor must be between-subjects. For example, if the research is investigating whether males or females are more adept at texting, the experiment probably involves entering text messages on a mobile phone. The independent variable is gender with two levels, male and female. The variable gender is between-subjects. Clearly, there is no choice. A participant cannot be male for half the testing, then female for the other half! Another example is handedness. Research investigating performance differences between left-handed and right-handed users requires a group of left-handed participants and a group of right-handed participants. Handedness, then, is a between-subjects factor. There is no choice.

Sometimes a factor must be within-subjects. The most obvious example is practice, since the acquisition of skill occurs within people, not between people. Practice is usually investigated by testing participants over multiple blocks of trials.

(a)

| Participant | Test Condition | | |
|---|---|---|---|
| 1 | A | B | C |
| 2 | A | B | C |

(b)

| Participant | Test Condition |
|---|---|
| 1 | A |
| 2 | A |
| 3 | B |
| 4 | B |
| 5 | C |
| 6 | C |

**FIGURE 5.6**

Assigning test conditions to participants: (a) Within-subjects. (b) Between-subjects.

For such designs, block is an independent variable, or factor, and there are multiple levels, such as block 1, block 2, block 3, and so on. Clearly, block is within-subjects since each participant is exposed to multiple blocks of testing. There is no choice.

Sometimes there is a choice. An important trade-off was noted above. That is, a within-subjects design requires fewer participants but requires more testing for each participant. There is a significant advantage to using fewer participants, since recruiting, scheduling, briefing, demonstrating, practicing, and so on is easier if there are fewer participants.

Another advantage of a within-subjects design is that the variance due to participants' predispositions will be approximately the same across test conditions. Predisposition, here, refers to any aspect of a participant's personality, mental condition, or physical condition that might influence performance. In other words, a participant who is predisposed to be meticulous (or sloppy!) is likely to carry their disposition in the same manner across the test conditions. For a between-subjects design, there are more participants and, therefore, more variability due to inherent differences between participants.

Yet another advantage of within-subjects designs is that it is not necessary to balance groups of participants—because there is only one group! Between-subjects designs include a separate group of participants for each test condition. In this case, balancing is needed to ensure the groups are more or less equal in terms of characteristics that might introduce bias in the measurements. Balancing is typically done through random assignment, but may also be done by explicitly placing participants in groups according to reasonable criteria (e.g., ensuring levels of computer experience are similar among the groups).

Because of the three advantages just cited, experiments in HCI tend to favor within-subjects designs over between-subjects designs.

However, there is an advantage to a between-subjects design. Between-subjects designs avoid interference between test conditions. Interference, here, refers to conflict that arises when a participant is exposed to one test condition and then switches to another test condition. As an example, consider an experiment that seeks to measure touch-typing speed with two keyboards. The motor skill acquired while learning to touch type with one keyboard is likely to adversely affect touch-typing with the other keyboard. Clearly, participants cannot "unlearn" one condition before testing on another condition. A between-subjects design avoids this because each participant is tested on one, and only one, of the keyboards. If the interference is likely to be minimal, or if it can be mitigated with a few warm-up trials, then the benefit of a between-subjects design is diminished and a within-subjects design is the best choice. In fact the majority of factors that appear in HCI experiments are like this, so levels of factors tend to be assigned within-subjects. I will say more about interference in the next section.

It is worth noting that in many areas of research, within-subjects designs are rarely used. Research testing new drugs, for example, would not use a within-subjects design because of the potential for interference effects. Between-subjects designs are typically used.

For an experiment with two factors it is possible to assign the levels of one factor within-subjects and the levels of the other factor between-subjects. This is a *mixed design*. Consider the example of an experiment seeking to compare learning of a text entry method between left-handed and right-handed users. The experiment has two factors: Block is within-subjects with perhaps 10 levels (block 1, block 2 … block 10) and handedness is between-subjects with two levels (left, right).

## 5.11 Order effects, counterbalancing, and latin squares

When the levels of a factor (test conditions) are assigned within-subjects, participants are tested with one condition, then another condition, and so on. In such cases, interference between the test conditions may result due to the order of testing, as noted above. In most within-subjects designs, it is possible—in fact, likely—that participants' performance will improve as they progress from one test condition to the next. Thus participants may perform better on the second condition simply because they benefited from practice on the first. They become familiar with the apparatus and procedure, and they are learning to do the task more effectively. Practice, then, is a confounding variable, because the amount of practice increases systematically from one condition to the next. This is referred to as a *practice effect* or a *learning effect*. Although less common in HCI experiments, it is also possible that performance will worsen on conditions that follow other conditions. This may follow from mental or physical fatigue—a *fatigue effect*. In a general sense, then, the phenomenon is an *order effect* or *sequence effect* and may surface either as improved performance or degraded performance, depending on the nature of the task, the inherent properties of the test conditions, and the order of testing conditions in a within-subjects design.

If the goal of the experiment is to compare the test conditions to determine which is better (in terms of performance on a dependent variable), then the confounding influence of practice seriously compromises the comparison. The most common method of compensating for an order effect is to divide participants into groups and administer the conditions in a different order for each group. The compensatory ordering of test conditions to offset practice effects is called *counterbalancing*.

In the simplest case of a factor with two levels, say, A and B, participants are divided into two groups. If there are 12 participants overall, then Group 1 has 6 participants and Group 2 has 6 participants. Group 1 is tested first on condition A, then on condition B. Group 2 is given the test conditions in the reverse order. This is the simplest case of a *Latin square*. In general, a Latin square is an $n \times n$ table filled with $n$ different symbols (e.g., A, B, C, and so on) positioned such that each symbol occurs exactly once in each row and each column.[9] Some examples of Latin square tables are shown in Figure 5.7. Look carefully and the pattern is easily seen.

[9]The name "Latin" in Latin square refers to the habit of Swiss mathematician Leonhard Euler (1707–1783), who used Latin symbols in exploring the properties of multiplication tables.

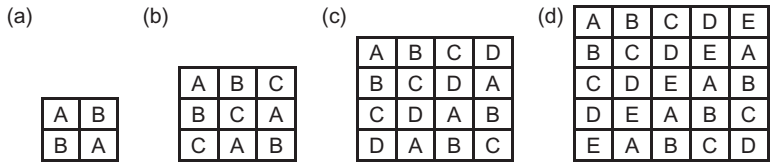



**FIGURE 5.7**

Latin squares: (a) 2 × 2. (b) 3 × 3. (c) 4 × 4. (d) 5 × 5.

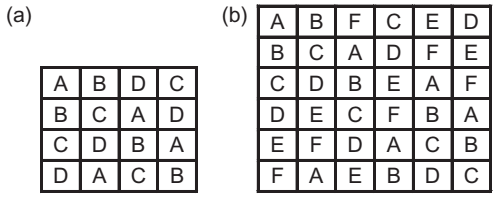



**FIGURE 5.8**

Balanced Latin squares where each condition precedes and follows other conditions an equal number of times: (a) 4 × 4. (b) 6 × 6.

The first column is in order, starting at A. Entries in the rows are in order, with wrap around.

A deficiency in Latin squares of order 3 and higher is that conditions precede and follow other conditions an unequal number of times. In the 4 × 4 Latin square, for example, B follows A three times, but A follows B only once. Thus an A-B sequence effect, if present, is not fully compensated for. A solution to this is a balanced Latin square, which can be constructed for even-order tables. Figure 5.8 shows 4 × 4 and 6 × 6 balanced Latin squares. The pattern is a bit peculiar. The first column is in order, starting at A. The top row has the sequence, A, B, *n*, C, *n*−1, D, *n*−2, etc. Entries in the second and subsequent columns are in order, with wrap around.

When designing a within-subjects counterbalanced experiment, the number of levels of the factor must divide equally into the number of participants. If a factor has three levels, then the experiment requires multiple-of-3 participants; for example, 9, 12, or 15 participants. If there are 12 participants, then there are three groups with 4 participants per group. The conditions are assigned to Group 1 in order ABC, to Group 2 in order BCA, and to Group 3 in order CAB (see Figure 5.7b). Let's explore this design with a hypothetical example.

An experimenter seeks to determine if three editing methods (A, B, C) differ in the time required for common editing tasks. For the evaluation, the following task is used:[10]

> Replace one 5-letter word with another, starting one line away.

[10]This is the same as task T1 described by Card, Moran, and Newell in an experiment to validate the keystroke-level model (KLM) (Card et al., 1980).

| Participant | Test Condition | | | Group | *Mean* | *SD* |
|---|---|---|---|---|---|---|
| | A | B | C | | | |
| 1 | 12.98 | 16.91 | 12.19 | 1 | 14.7 | 1.84 |
| 2 | 14.84 | 16.03 | 14.01 | | | |
| 3 | 16.74 | 15.15 | 15.19 | | | |
| 4 | 16.59 | 14.43 | 11.12 | | | |
| 5 | 18.37 | 13.16 | 10.72 | 2 | 14.6 | 2.46 |
| 6 | 15.17 | 13.09 | 12.83 | | | |
| 7 | 14.68 | 17.66 | 15.26 | | | |
| 8 | 16.01 | 17.04 | 11.14 | | | |
| 9 | 14.83 | 12.89 | 14.37 | 3 | 14.4 | 1.88 |
| 10 | 14.37 | 13.98 | 12.91 | | | |
| 11 | 14.40 | 19.12 | 11.59 | | | |
| 12 | 13.70 | 16.17 | 14.31 | | | |
| *Mean* | 15.2 | 15.5 | 13.0 | | | |
| *SD* | 1.48 | 2.01 | 1.63 | | | |

**FIGURE 5.9**

Hypothetical data for an experiment with one within-subjects factor having three levels (A, B, C). Values are the mean task completion time(s) for five repetitions of an editing task.

The following three editing methods are compared (descriptions are approximate):

Method A: arrow keys, BACKSPACE, type
Method B: search and replace dialog
Method C: point and double click with the mouse, type

Twelve participants are recruited. To counterbalance for learning effects, participants are divided into three groups with the tasks administered according to a Latin square (see Figure 5.7b). Each participant does the task five times with one editing method, then again with the second editing method, then again with the third. The mean task completion time for each participant using each editing method is tabulated. (See Figure 5.9.) Overall means and standard deviations are also shown for each editing method and for each group. Note that the left-to-right order of the test conditions in the figure applies only to Group 1. The order for Group 2 was BCA and for Group 3 CAB (see Figure 5.7b).

At 13.0 s, the mouse method (C) was fastest. The arrow-key method (A) was 17.4 percent slower at 15.2 s, while the search-and-replace method (B) was 19.3 percent slower at 15.5 s. (Testing for statistical significance in the differences is discussed in the next chapter.) Evidently, counterbalancing worked, as the group means are very close, within 0.3 s. The tabulated data in Figure 5.9 are not typically provided in a research report. More likely, the results are presented in a chart, similar to that in Figure 5.10.

Although counterbalancing worked in the above hypothetical example, there is a potential problem for the $3 \times 3$ Latin square. Note in Figure 5.7b that B follows A twice, but A follows B only once. So there is an imbalance. This cannot be avoided in Latin squares with an odd number of conditions. One solution in this case is to counterbalance by using all sequences. The $3 \times 3$ case is shown in Figure 5.11.

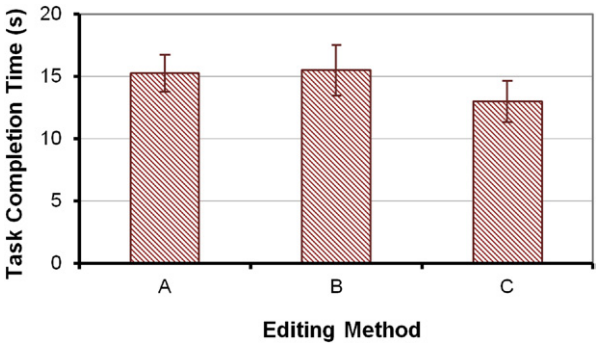



**FIGURE 5.10**

Task completion time(s) by editing method for the data in Figure 5.9. Error bars show ±1 *SD*.

| | | |
|---|---|---|
| A | B | C |
| A | C | B |
| B | C | A |
| B | A | C |
| C | A | B |
| C | B | A |

**FIGURE 5.11**

Counterbalancing an odd number of conditions using all (*n*!) combinations.

There are 3! = 6 combinations. Balancing is complete (e.g., B follows A three times, A follows B three times). MacKenzie and Isokoski (2008) used such an arrangement in an experiment with 18 participants, assigning 3 participants to each order.

Yet another way to offset learning effects is to randomize the order of conditions. This is most appropriate where (a) the task is very brief, (b) there are many repetitions of the task, and (c) there are many test conditions. For example, experiments that use point-select tasks often include movement direction, movement distance, or target size as factors (Figure 5.12).

The test conditions in Figure 5.12 might appear as factors in an experiment even though the experiment is primarily directed at something else. For example, research comparing the performance of different pointing devices might include device as a factor with, say, three levels (mouse, trackball, stylus). Movement direction, movement distance, and target size might be varied to ensure the tasks cover a typical range of conditions. Treating these conditions as factors ensures they are handled in a systematic manner. To ensure equal treatment, the conditions are chosen at random without replacement. Once all conditions have been used, the process may repeat if multiple blocks of trials are desired.

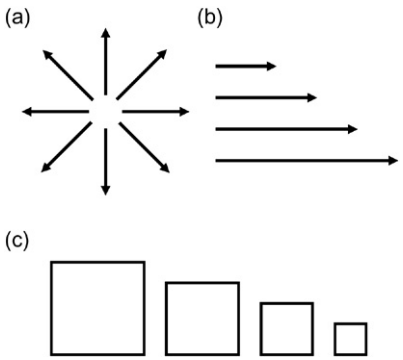



**FIGURE 5.12**

Test conditions suitable for random assignment: (a) Movement direction. (b) Movement distance. (c) Target size.

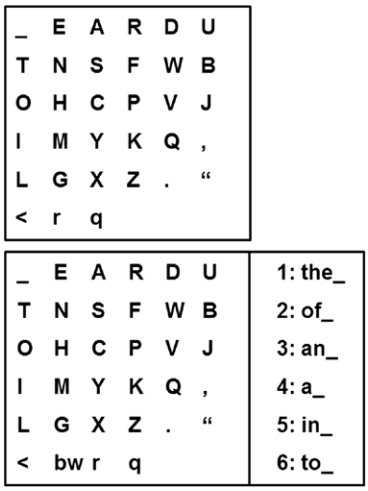
## 5.12 Group effects and asymmetric skill transfer

If the learning effect is the same from condition to condition in a within-subjects design, then the group means on a dependent variable should be approximately equal.[11] This was demonstrated above (see Figure 5.9). In other words, the advantage due to practice for a condition tested later in the experiment is offset equally by the disadvantage when the same condition is tested earlier in the experiment. That's the point of counterbalancing. However, there are occasions where different effects appear for one order (e.g., A→B) compared to another (e.g., B→A). In such cases there may be a *group effect*—differences across groups in the mean scores on a dependent variable. When this occurs, it is a problem. In essence, counterbalancing did not work. A group effect is typically due to *asymmetric skill transfer*—differences in the amount of improvement, depending on the order of testing.

We could develop an example of asymmetric skill transfer with hypothetical data, as with the counterbalancing example above; however, there is an example data set in a research report where an asymmetric transfer effect is evident. The example provides a nice visualization of the effect, plus an opportunity to understand why asymmetric skill transfer occurs. So we'll use that data. The experiment compared two types of scanning keyboards for text entry (Koester and Levine, 1994a). Scanning keyboards use an on-screen virtual keyboard and a single key or switch for input. Rows of keys are highlighted one by one (scanned). When the row bearing the desired letter is highlighted, it is selected. Scanning enters the row and advances left to right. When the key bearing the desired letter is highlighted it is selected and the letter is added to the text message. Scanning keyboards provided a convenient text entry method for many users with a physical disability.

[11] There is likely some difference, but the difference should not be statistically significant.

(a)




(b)

| Testing Half | | Group |
|---|---|---|
| First (Trials 1-10) | Second (Trials 11-20) | |
| **20.42** | 27.12 | 1 |
| **22.68** | 28.39 | |
| **23.41** | 32.50 | |
| **25.22** | 32.12 | |
| **26.62** | 35.94 | |
| **28.82** | 37.66 | |
| **30.38** | 39.07 | |
| **31.66** | 35.64 | |
| **32.11** | 42.76 | |
| **34.31** | 41.06 | |
| 19.47 | **24.97** | 2 |
| 19.42 | **27.27** | |
| 22.05 | **29.34** | |
| 23.03 | **31.45** | |
| 24.82 | **33.46** | |
| 26.53 | **33.08** | |
| 28.59 | **34.30** | |
| 26.78 | **35.82** | |
| 31.09 | **36.57** | |
| 31.07 | **37.43** | |

**FIGURE 5.13**

Experiment comparing two scanning keyboards: (a) Letters-only keyboard (LO, *top*) and letters plus word prediction keyboard (L + WP, *bottom*). (b) Results for entry speed in characters per minute (cpm). Shaded cells are for the LO keyboard.

The experiment compared a letters-only (LO) scanning keyboard with a similar keyboard that added word prediction (L + WP). The keyboards are shown in Figure 5.13a. Six participants entered 20 phrases of text, 10 with one keyboard, followed by 10 with the other. To compensate for learning effects, counterbalancing was used. Participants were divided into two groups. Group 1 entering text with the LO keyboard first, then with the L + WP keyboard. Group 2 used the keyboards in the reverse order. Although not usually provided in a report, the results were given in a table showing the entry speed in characters per minute (cpm). The data are reproduced in Figure 5.13b as they appeared in the original report (Koester and Levine, 1994a, Table 2). The two columns show the sequence of testing: first half, then second half. The shaded and un-shaded cells show the results for the LO and L + WP keyboards respectively, thus revealing the counterbalanced order.

There are at least three ways to summarize the data in Figure 5.13b. The overall result showing the difference between the LO and L + WP keyboards is shown in the left-side chart in Figure 5.14. Clearly, there was very little difference between the two keyboards: 30.0 cpm for the LO keyboard versus 30.3 cpm for the L + WP keyboard. The L + WP keyboard was just 1 percent faster. The error bars are large, mostly due to the improvement from trial to trial, as seen in Figure 5.13b.

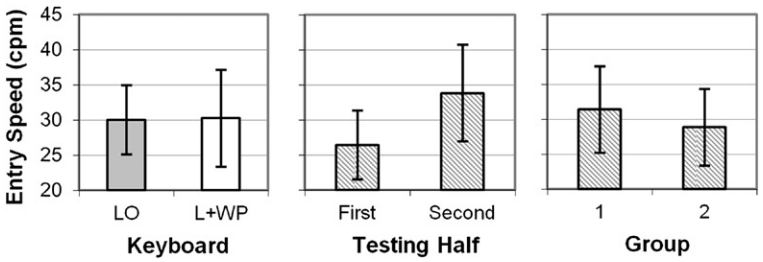



**FIGURE 5.14**

Three ways to summarize the results in Figure 5.13b, by keyboard (*left*), by testing half (*center*), and by group (*right*). Error bars show ±1 *SD*.

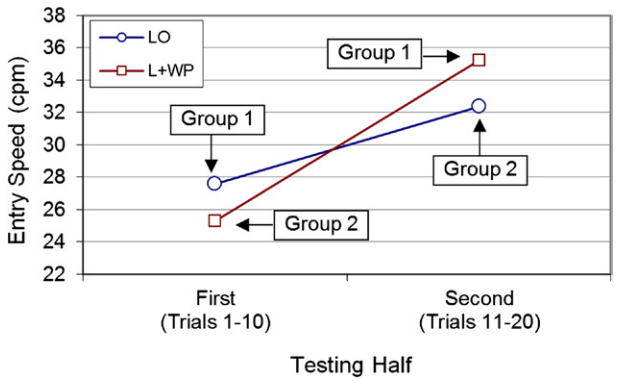



**FIGURE 5.15**

Demonstration of asymmetric skill transfer. The chart uses the data in Figure 5.13b.

The center chart in Figure 5.14 shows another view of the results, comparing the first half and second half of testing. A learning effect is clearly seen. The overall entry speed was 26.4 cpm in the first half of testing (trials 1 to 10) and 33.8 cpm, or 28 percent higher, in the second half of testing (trials 11 to 20). Learning is fully expected, so this result is not surprising.

Now consider the right-side chart in Figure 5.14. Counterbalancing only works if the *order effects* are the same or similar. This implies that the performance benefit of an LO→L + WP order is the same as the performance benefit of an L + WP→LO order. If so, the group means will be approximately equal. (This was demonstrated earlier in the counterbalancing example; see Figure 5.9). The right-side chart in Figure 5.14 reveals a different story. The mean for Group 1 was 31.4 cpm. The mean for Group 2 was lower at 28.8 cpm. For some reason, there was an 8 percent performance disadvantage for Group 2. This is an example of asymmetric skill transfer. Figure 5.15 illustrates. The figure reduces the data in Figure 5.13b to four points, one for each quadrant of 10 trials. Asymmetry is clearly seen in the cross-over of

the lines connecting the LO points and L + WP points between the first half and second half of testing.

If counterbalancing had worked, the lines in Figure 5.15 would be approximately parallel. They are not parallel because of the asymmetry in the LO→L + WP order versus the L + WP→LO order. Asymmetric skill transfer is usually explainable by considering the test conditions or the experimental procedure. For this experiment, the effect occurs because of the inherent differences in entering text with the letters-only (LO) keyboard versus entering text with the letters plus word prediction (L + WP) keyboard. In fact, this example provides an excellent opportunity to understand why asymmetric skill transfer sometimes occurs. Here is the explanation. The L + WP keyboard is an enhanced version of the LO keyboard. The basic method of entering letters is the same with both keyboards; however, the L + WP keyboard adds word-prediction, allowing words to be entered before all letters in the word are entered. It is very likely that entering text first with the LO keyboard served as excellent practice for the more difficult subsequent task of entering text with the L + WP keyboard. To appreciate this, examine the two points labeled Group 1 in Figure 5.15. Group 1 participants performed better overall because they were tested initially with the easier LO keyboard before moving on the enhanced L + WP keyboard. Group 2 participants fared less well because they were tested initially on the more difficult L + WP keyboard.

The simplest way to avoid asymmetric skill transfer is to use a between-subjects design. Clearly, if participants are exposed to only one test condition, they cannot experience skill transfer from another test condition. There are other possibilities, such as having participants practice on a condition prior to data collection. The practice trials seek to overcome the benefit of practice in the earlier condition, so that the measured performance accurately reflects the inherent properties of the test condition. It is not clear that this would work in the example. Participants cannot "unlearn."

In the end, the performance difference between the LO and L + WP keyboards remains an outstanding research question. The practice effect (28%) was much greater than the group effect (8%), so it is difficult to say whether word prediction in the L + WP keyboard offers a performance advantage. Clearly, there is a *benefit* with the L + WP keyboard, because words can be entered before all the letters are entered. However, there is also a *cost*, since users must attend to the on-going prediction process, and this slows entry. To determine whether the costs outweigh the benefits in the long run, a longitudinal study is required. This is examined in the next section.

## 5.13 Longitudinal studies

The preceding discussion focused on the confounding influence of learning in experiments where an independent variable is assigned within-subjects. Learning effects—more generally, order effects—are problematic and must be accommodated in some way, such as counterbalancing. However, sometimes the research

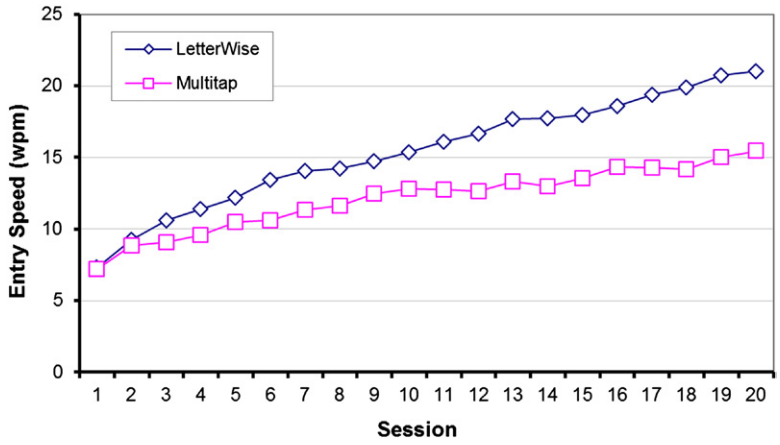



**FIGURE 5.16**

Example of a longitudinal study. Two text entry methods were tested and compared over 20 sessions of input. Each session involved about 30 minutes of text entry.

has a particular interest in learning, or the acquisition of skill. In this case, the experimental procedure involves testing users over a prolonged period while their improvement in performance is measured. Instead of eliminating learning, the research seeks to observe it and measure it. An experimental evaluation where participants practice over a prolonged period is called a *longitudinal study*.

In a longitudinal study, "amount of practice" is an independent variable. Participants perform the task over multiple units of testing while their improvement with practice is observed and measured. Each unit of testing is a level of the independent variable. Various names are used for the independent variable, but a typical example is *Session* with levels Session 1, Session 2, Session 3, and so on. An example is an experiment comparing two text entry methods for mobile phones: multi-tap and *LetterWise* (MacKenzie, Kober, Smith, Jones, and Skepner, 2001). For English text entry, LetterWise requires an average of 44 percent fewer keystrokes than does multi-tap. However, a performance benefit might not appear immediately, since users must learn the technique. Furthermore, learning occurs with both methods, as participants become familiar with the experimental procedure and task. However, it was felt that the reduction in keystrokes with LetterWise would eventually produce higher text entry speeds. To test this, a longitudinal study was conducted, with entry method assigned between-subjects. The results are shown in Figure 5.16. Indeed, the conjectured improvement with practice was observed. Initial entry speeds were about 7.3 wpm for both methods in Session 1. With practice, both methods improved; however, the improvement was greater with LetterWise because of the ability to produce English text with fewer keystrokes on average. By Session 20, text entry speed with LetterWise was 21.0 wpm, about 36 percent higher than the rate of 15.5 wpm for multi-tap.

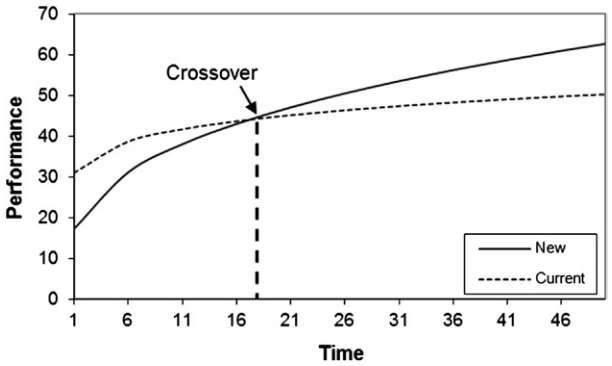



**FIGURE 5.17**

Crossover point. With practice, human performance with a new interaction technique may eventually exceed human performance using a current technique.

*(From MacKenzie and Zhang, 1999)*

Performance trends in longitudinal studies, as shown in Figure 5.15, are often accompanied by an equation and best-fitting curve demonstrating the *power law of learning*. Examples are given in Chapter 7, section 7.2.5 (Skill Acquisition).

In many situations, the goal of a longitudinal study is to compare the viability of a new technique against current practice. Here, current practice is any conventional interaction that is quantifiable using a performance measure. Examples include text entry, editing, pointing, selecting, searching, panning, zooming, rotating, drawing, scrolling, menu access, and so on. If users are experienced with a current interaction technique, then relatively poorer initial performance is expected with the new technique. But as learning progresses, the performance trends may eventually cross over, wherein performance with the new technique exceeds that with current practice. This is illustrated in Figure 5.17.

As an example, consider the ubiquitous Qwerty keyboard. Although improved designs have been proposed, users experienced with a Qwerty keyboard are unlikely to demonstrate an immediate improvement in performance with an alternative design. Considerable practice may be required before performance on the new keyboard exceeds that with the Qwerty keyboard. The Dvorak simplified keyboard (DSK), for example, has been demonstrated in longitudinal studies to provide a speed advantage over Qwerty (see Noyes, 1983 for a review). Yet Qwerty remains the dominant form factor for computer keyboards. From a practical standpoint, learning a new technique bears a *cost*, since performance is initially superior with the current technique. However, after the crossover point is reached, the new technique provides a *benefit*, since performance is superior compared to current practice. The cost-benefit trade-off is shown in Figure 5.18.

Despite the long-term benefits evident in Figure 5.18, new technologies often languish in the margins while established but less-optimal designs continue to dominate the marketplace. Evidently, the benefits are often insufficient to overcome the costs.

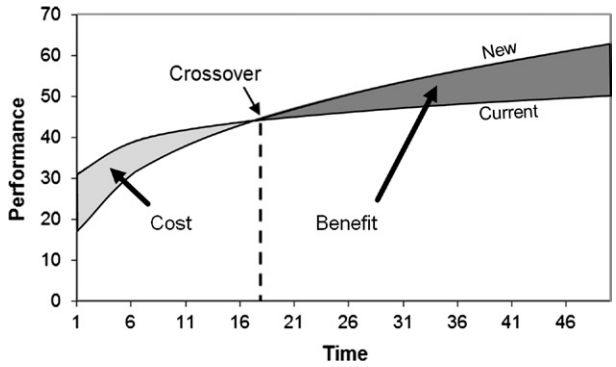



**FIGURE 5.18**

Cost-benefit progression in learning a new interaction technique where there is existing skill with current practice.

With respect to the Qwerty debate, there are two such costs. One is the cost of manufacturing and retooling. Keyboards are electro-mechanical devices, so new designs require ground-up reengineering and new manufacturing materials and procedures. This is expensive. The other cost lies in overcoming user perceptions and attitudes. By and large, users are change-adverse: they are reluctant to give up habits they have acquired and are comfortable with. Simply put, users are "unwilling to change to a new keyboard layout because of the retraining required" (Noyes, 1983, p. 278).

One interesting example is a soft or virtual keyboard, as commonly used on touchscreen phones or personal digital assistants (PDAs). Input is typically with a finger or stylus. Most such keyboards use the Qwerty letter arrangement. However, since the keyboard is created in software, there is no retooling cost associated with an alternative design. Thus there is arguably a better chance for an optimized design to enter the marketplace. One idea to increase text entry speed is to rearrange letters with common letters clustered near the center of the layout and less common letters pushed to the perimeter. The increase in speed results from the reduction in finger or stylus movement. However, since users are unfamiliar with the optimized letter arrangement, performance is initially poor (while they get accustomed to the letter arrangement). If learning the new technique is likely to take several hours or more, then the evaluation requires a longitudinal study, where users are tested over multiple sessions of input. Eventually, the crossover point may appear. This idea is explored further in Chapter 7, section 7.2.5 (Skill Acquisition).

## 5.14 Running the experiment

When the experiment is designed, the apparatus built and tested, the participants recruited and scheduled, then testing begins. But wait! Are you sure the time to begin has arrived? It is always useful to have a pilot test (yes, one more pilot test)

with one or two participants. This will help smooth out the protocol for briefing and preparing the participants. It will serve as a check on the amount of time needed for each participant. If the testing is scheduled for one hour, it is important that all the testing combined with briefing, practicing, etc., comfortably fit into one hour. A final tweak to the protocol may be necessary. Better now than to have regrets later on.

So the experiment begins. The experimenter greets each participant, introduces the experiment, and usually asks the participants to sign consent forms. Often, a brief questionnaire is administered to gather demographic data and information on the participants' related experience. This should take just a few minutes. The apparatus is revealed, the task explained and demonstrated. Practice trials are allowed, as appropriate.

An important aspect of the experiment is the instructions given to participants. Of course, the instructions depend on the nature of the experiment and the task. For most interaction tasks, the participant is expected to proceed quickly and accurately. These terms—quickly and accurately—are subject to interpretation, as well as to the capabilities of participants. What is reasonably quick for one participant may be unattainable by another. Performing tasks reasonably quick and with high accuracy, but at a rate comfortable to the individual, is usually the goal. Whatever the case, the instructions must be carefully considered and must be given to all participants in the same manner. If a participant asks for clarification, caution must be exercised in elaborating on the instructions. Any additional explanation that might motivate a participant to act differently from other participants is to be avoided.

The experimenter plays a special role as the public face of the experiment. It is important that the experimenter portrays himself or herself as neutral. Participants should not feel they are under pressure to produce a specific outcome. Deliberately attempting to perform better on one test condition compared to another is to be avoided. Also, participants should not sense a particular attitude in the experimenter. An overly attentive experimenter may make the participant nervous. Similarly, the experimenter should avoid conveying indifference or disregard. If the experimenter conveys a sense of not caring, the participant may very well act with little regard to his or her performance. A neutral manner is preferred.

## STUDENT EXERCISES

**5-1.** It was noted above that independent variables in research papers are sometimes identified without being given a name. Review some experimental research papers in HCI and find three examples of this. Propose a name for the independent variable and give examples of how to improve the paper, properly identifying both the name of the variable and the levels of the variable. Examine how the independent variables and the levels of the independent variables (test conditions) were referred to in the paper. Point out any inconsistencies.

**5-2.** Review some HCI experimental research papers and find three examples where an independent variable was assigned between-subjects. Briefly describe the rationale.

**5-3.** Find an example of an HCI paper describing an experiment that included three (or more) independent variables. Construct a chart similar to Figure 5.2, labeling the independent variables. Indicate which effects were studied and which effects (if any) were not studied.

**5-4.** Users and computers (Part I). Design and administer a simple questionnaire to about 25 computer users. Solicit four items of information: gender, age, hours per day using a computer, and preferred brand of computer (Mac versus PC). Use a ratio-scale questionnaire item to solicit the respondent's age. Write a brief report on the findings. This exercise continues in Chapter 6 (Part II) and Chapter 7 (Part III).

**5-5.** Find an example of an HCI paper describing a longitudinal study. (Preferably, find an example not cited in this book.) Write a brief report describing the testing procedure. Over what period of time did testing extend, and how were the trials organized? Identify the main independent variable (and levels) and discuss how it was administered. Note features such as counterbalancing, within-subjects or between-subjects assignment, and whether practice trials were administered.