College of Information Sciences and Technology
The Pennsylvania State University

How to Run Experiments: A Practical Guide to Research with Human Participants

Frank E. Ritter[1], Jong W. Kim[2], Jonathan H. Morgan[1], and Richard A. Carlson[1]
frank.ritter@psu.edu jong.kim@ucf.edu jhm5001 @psu.edu racarlson@psu.edu

The Pennsylvania State University[1]
University Park, PA 16802

University of Central Florida[2]
Orlando, FL 32816

Technical Report No. ACS 2012-01

13 February 2012

Phone +1 (814) 865-4453 Fax +1 (814) 865-5604

College of IST, IST Building, University Park, PA 16802

Abstract and Preface

There are few practical guides on how to prepare and run experiments with human participants in a laboratory setting. In our experience, we have found that students are taught how to design experiments, and how to analyze data in courses such as *Design of Experiments and Statistics*. On the other hand, the dearth of materials available to students preparing and running experiments has often led to a gap between theory and practice in this area, which is particularly acute outside of psychology departments. Consequently, labs frequently must not only impart these practical skills to students informally, but must also address misunderstandings arising from this divorce of theory and practice in their formal education.

This guide presents advice that can help young experimenters and research assistants to run experiments effectively and more comfortably with human participants. In this book, our purpose is to provide hands-on knowledge about and actual procedures for experiments. We hope this book will help students of psychology, engineering, and the sciences to run studies with human participants in a laboratory setting. This will particularly help students (or instructors and researchers) who are not in large departments, or are running participants in departments that do not have a large or long history of experimental studies of human behavior. This book is also intended to help people who are starting to run user and usability studies in industry.

We are generally speaking here from our background running cognitive psychology, cognitive ergonomics, and human-computer interaction studies. Because it is practical advice, we do not cover experimental design or data analyses. This practical advice will be less applicable in more distant areas, or when working in more complex situations, but may be still of use. For example, we do not cover how to use complex machinery, such as fMRI or ERP. We also do not cover field studies or studies that in the US require a full IRB review. This means that we do not cover how to work with unusual populations such as prisoners, animals, and children, or how to take and use measures that include risks to the subjects or to the experimenter (e.g., saliva, blood samples, or private information).

This book arose during a discussion at Jong Kim's PhD graduation. Ritter asked Kim what he thought were places where more training might have been helpful; the conversation turned to experimental methods and the tactics and details of running studies. During the graduation ceremony, they outlined this book—a worthy genesis for a book we think. Since then, we and others have used it to teach both in classrooms and at conference tutorials, and it has been expanded, corrected, and extended.

We have addressed this book to advanced undergraduates and early graduate students starting to run experiments without previous experience, but we believe this guide will be useful to anyone who is starting to run research studies, training people to run studies, or studying the experimental process. It should also be useful to researchers in industry who are also starting to run studies.

When running an experiment, insuring its repeatability is of greatest importance—it is critical to address variations in either method or in participant behavior. Running an experiment in exactly the same way regardless of who is conducting it or where (e.g., different research teams or laboratories) is essential. In addition, reducing unanticipated variance in the participants' behavior is key to an experiment's repeatability. This book will help you achieve these requirements, increasing both your comfort and that of the participants who participate in your experiments. We hope you find it relevant and useful.

This book consists of several sections with multiple appendices. As an advance organizer we briefly describe each section's contents.

Chapter 1, Overview of the Research Process, describes where experiments fit into the research process. If you have taken either an experimental methods course or a research design course, you can skip this chapter. If, on the other hand, you are either a new research assistant or are working on a project in which you are unclear of your role or how to proceed, this chapter may provide some helpful context. This chapter also introduces several running examples.

Chapter 2, Preparation for Running Experiments, describes pertinent topics for preparing to run your experiment—such as supplemental reading materials, recruitment of participants, choosing experimental measures, and getting Institutional Review Board (IRB) approval for experiments involving participants.

Chapter 3, Potential Ethical Problems, describes ethical considerations necessary for safely running experiments with human participants—i.e., how to ethically recruit participants, how to handle data gathered from participants, how to use that data, and how to report that data. Being vigilant and aware of these topics is an important component to rigorous, as well as ethical, research.

Chapter 4, Risks to Validity to Avoid While Running an Experiment, describes risks that can invalidate your experimental data. If you fail to avoid these types of risks, you may obtain either false or uninterruptible results from your experiment. Thus, before starting your study, you should be aware of these risks and how to avoid them.

Chapter 5, Running a Research Study, describes practical information about what you have to do when you run experiments. This section will give an example procedure that you can follow.

Chapter 6, Concluding a Research Session and Study, describes practical information about what to do at the conclusion of each experimental session and at the end of a study.

Chapter 7, Afterward, summarizes the book and describes the appendices.

The Appendixes include an example checklist for starting a study, a checklist for setting up a study, an example consent form, an example debriefing form, and an example IRB form. The details and format of these forms will vary by lab and IRB committee, but the materials in the appendixes provide examples of the style and tone. There is also an appendix on how this material could apply to online studies.

A web site holding supplementary material is available a http://acs.ist.psu.edu/how-to-run-experiments

Acknowledgements

Preparation of this manuscript was partially sponsored by a grant from the Division of Human Performance Training, and Education at the Office of Naval Research, under Contracts # W911QY-07-01-0004 and N00014-11-1-0275. The views and conclusions contained in this report are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government or the Pennsylvania State University.

Christine Cardone at Sage provided some encouragement when we needed it. Numerous people have given useful comments, and when they have used it in teaching we note that as well here. Ray Adams (Middlesex), Susanne Bahr (Florida Institute of Technology), Ellen Bass, Gordon Baxter (St. Andrews), Stephen Broomell, Karen Feigh (Georgia Institute of Technology, several times), Katherine Hamilton, William (Bill) Kennedy, Alex Kirlik (U. of Illinois), Michelle Moon, Razvan Orendovici, Erika Poole, Michael (Q) Qin (NSMRL/U. of Connecticut), Joseph Sanford, Robert West (Carleton), Hongbin Wong (U. of Texas/Houston), Kuo-Chuan (Martin) Yeh, Xiaolong (Luke) Zhang (PSU), and several anonymous reviewers have provided useful comments. Ryan Moser and Joseph Sanford have helped prepare this manuscript, but any incompleteness and inadequacies remain the fault of the authors.

Blurbs

I should've read this 2 years ago. I think this will be really helpful and informative not only for grad students who are new in the field but also for researchers who have some experiences but are not sure if they are doing the right things or if there is any better way to do things.
--- Michelle Moon, CMU graduate student

“This is exactly what I need.” Robert St. Amant, 8 May 2010

“This is just common sense.” Psychology professor, summer 2011.

Table of Contents

1 Overview of the Research Process

Individuals who conduct behavioral research with human participants as part of their careers, like other specialists, have developed a set of good practices, standard methodology, and specialized vocabulary for discussing the research process. If you have taken a course in research methods, or read a standard research methods textbook, much of this vocabulary will be familiar to you. We assume, however, that many readers of this book are new to research or will find some reminders useful. If you are new, the good practical practices learned through a hands-on apprenticeship might not be available to you in your situation, and that is the purpose of this book.

We focus here on behavioral research, by which we mean research with the primary object of observing, understanding, and predicting actions. These actions can be primarily physical actions, but typically behavioral research is concerned with the meaning of behavior—the answers communicated by speech, key presses, or other means, the effect of actions in achieving goals, and so on. Behavioral research is often contrasted with neuroscience research, which is primarily concerned with understanding how the brain and nervous system support behavior. Much of what we have to say is drawn from the field of experimental psychology, but researchers in many fields make use of behavioral research methods.

1.1 Overview

This book is primarily about conducting *experiments*. In everyday usage, "experimenting" simply means trying something out—a new recipe, a different word-processing program, or perhaps a new exercise routine. In this book, however, *experiment* has a more formal definition. There are two primary elements to that definition. First, we are concerned with ***control***—not controlling our participants, though we'll sometimes want to do some of that—but with controlling the circumstances under which we make our observations. Second, we are concerned with ***cause and effect***, or the relationship between an *independent variable* and a *dependent variable.* Winston (1990) traces this use of the term "experiment" to Woodworth's (1938) classic text, which is perhaps the most influential book on methodology in the history of experimental psychology.

The concept of control is important because, more or less, almost everything affects almost everything else in an experimental situation. For example, if we are interested in whether people like a new computer interface, we have to recognize that their reactions may be influenced by their mood, the time of day, extraneous noises, their familiarity with similar interfaces, and so on. Much of this book is about how the researcher can achieve control of the circumstances under which they make their observations. Sometimes this is done by actually controlling the circumstances—making our observations at consistent times of day or eliminating extraneous noises. Sometimes, it is done using statistical methods to account for factors we cannot literally control. These statistical methods are also referred to as *control*.

Sometimes, controlling the conditions of our observations is sufficient to answer a research question. If we simply want to know whether individuals find an interface pleasant to work with, we can ask them to use the interface under controlled conditions, and assess their reactions through interviews or rating schemes. This is called controlled observation. If controlled observation is sufficient for your research purpose, you will still find much useful advice in this book about how to achieve control. Much of product development uses these types of studies, and a lot can be learned about how people use technology in this way. More often, though, observations like this beg the question, "Compared to what?"

In contrast to controlled observation, an experiment—sometimes called a "true experiment"—involves manipulating an independent variable. For example, our question might not be whether individuals find an interface pleasant to work with, but whether they find Interface A more pleasant than Interface B, or vice versa. In this case, the independent variable would be the interface, and the variable would have two levels, A and B. The dependent variable would be whatever we measure—users' ratings, their success in using the interface, and so on. A true experiment has at least one independent and one dependent variable, but it is possible to have more of both. Independent variables are also sometimes called "factors."

It is important to know some of the other jargon common to a psychology laboratory. A *stimulus* is an environmental event—typically, now, a display on a computer screen—to which a subject *responds*. Most experiments involve numerous *trials*—individual episodes in which a stimulus is displayed and a response measured. Often, these trials are grouped into *blocks* to provide sets of observations that serve as units of analysis, or to mark events in the experimental procedure such as rest periods for subjects. The language of stimulus, response, trial, and block is often awkward for experiments using complex, dynamic tasks. Nevertheless, these terms are used frequently enough in the field and in this book that it should become familiar. The terms we have introduced here, and others that will be useful in reading this book, are briefly defined in Table 1.1.

We will also frequently mention "the literature" relevant to an experiment. This simply means the accumulated articles, chapters, and books that report related research. Reading relevant scientific literature is important because it can help sharpen your research question, allow you to avoid mistakes, and deepen your understanding of the results. University libraries generally have powerful database tools for searching for research related to a particular topic. Common databases are PsycInfo (a very complete database maintained by the American Psychological Association), and Web of Science (a database allowing citation searches, provided by Thomson Reuters). The trick to searching the literature using these databases is to know the appropriate keywords. For many researchers conducting experiments for the first time, this can be an obstacle. For example, understanding how individuals use an interface may, depending on the specific question, lead to issues relating to working memory, attention, perception, skill, or motor control. Each of these domains of cognitive research has generated a vast literature. Sometimes, reading a handful of abstracts found with a database search will help to focus the search. Another resource for many topics is online references like Wikipedia, but you should be aware that Wikipedia articles provide starting points and are not considered primary resources because they are not fully reviewed or archival, and change over time. Often, the best strategy is to find a friendly expert, tell him or her about your research question, and ask for suggestions on what to read. We will return to the topic of relevant literature in Chapter 2.

With this as background, we turn to an overview of the research process.

Table 1.1: Definitions

Block: A portion of an experiment distinguished by breaks for subjects, shifts in procedure, and so on. Typically, a block is a set of trials. Often, blocks serve as units of analysis.

Condition (experimental condition): A subset of the experiment defined by a level or value of an independent variable.

Control: The holding constant by procedure or statistical procedure of variables other than the independent variable.

Dependent variable (DV): A variable that depends on the subjects' behavior, such as the time to respond or the accuracy of the response.

Experiment: A study in which an independent variable is manipulated, a dependent variable measured, and other variables controlled.

Experimenter (E): A member of the research team who interacts directly with subjects. The experimenter may be one of the researchers, or someone whose sole role is interacting with subjects to carry out experimental procedures.

Hypothesis: A prediction about the outcome of an experiment, stated as an expected relationship between the independent and dependent variables.

Independent variable (IV): A variable that is manipulated by the researcher; the values of an independent variable are independent of the subjects' behavior.

Informed consent: The process by which subjects are first informed about what they will experience if they participate in the study; and second indicate whether they consent to take part.

Investigator, principal investigator (PI), researcher, lead researcher: The individuals responsible for making scientific decisions and judgments. "Principal investigator" refers to the individual who takes final responsibility for the study to granting agencies, the IRB, etc. "Lead researcher" refers to the individual who designs and makes scientific judgments about the study. In practice, although the principal investigator role is usually officially defined, the distinctions among roles may be blurred.

IRB: Institutional Review Board, the panel that is responsible for reviewing experimental procedures for compliance with ethical and regulatory standards.

Null hypothesis: The hypothesis in which the independent variable does not affect the dependent variable. The null hypothesis serves a special role in tests of statistical significance.

Response: The units of the subjects' behavior. Responses may be key presses, verbal answers, moves in a problem environment, and so on.

Statistical significance: A criterion for deciding that the experimental hypothesis is sufficiently more likely than the null hypothesis to allow the conclusion that the independent variable affects the dependent variable.

Statistical power: The ability or sensitivity of an experiment to detect an effect of an independent variable.

Stimulus: An environmental event to which a subject responds.

Subject or participant (S or P): An individual who performs the experimental task and whose behavior is the object of analysis.

Trial: An episode within an experiment in which a stimulus occurs and the subject responds.

1.2 Overview of the research process

Figure 1-1 summarizes the research process, with notes about where in the book the step is discussed. A glance at the figure shows that the process is iterative—rarely do even experienced researchers generate an experiment without pilot testing. The dashed lines show loops that sometimes occur, and the solid line shows a loop that nearly always occurs. Piloting nearly always results in the refinement of the initial procedure.

The figure also shows that the process generally involves others, both other colleagues in your lab and outside institutions. Further, research with human subjects conducted in a university or other organization that receives federal funding (in the United States) requires the approval of an Institutional Review Board (IRB). An IRB evaluates the experimental procedure for possible risks to the participants and other ethical ambiguities, such as potential conflicts of interest. Other organizations and countries often have similar requirements. We return to the issue of ethics and IRB review in Chapters 2 and 3; for now, the point is that planning an experiment almost always requires consultation with groups outside of the research team.

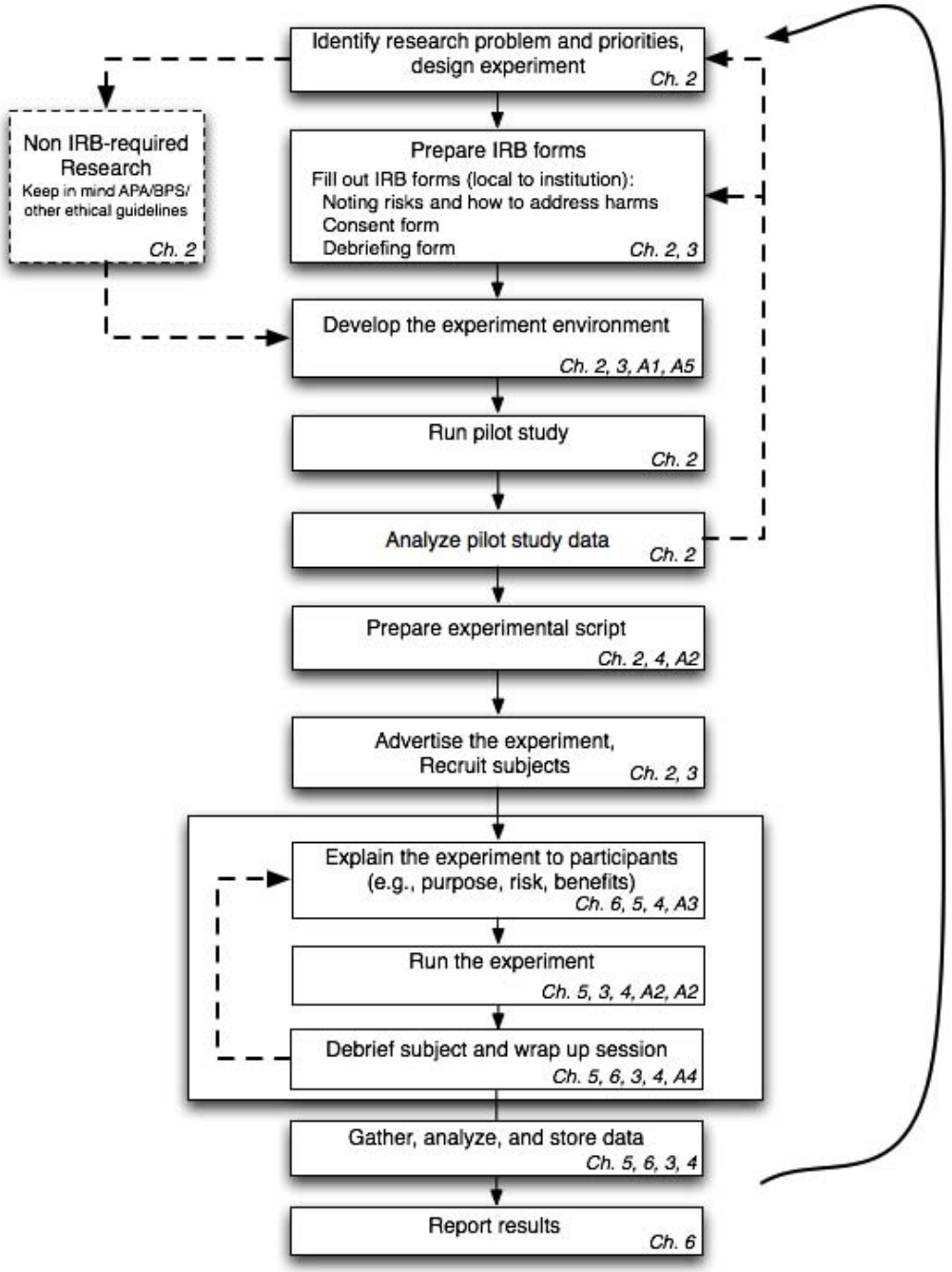


Figure 1-1. A pictorial summary of the research process. This is similar to, but developed separately from Bethel and Murphy's (2010) figure for human-robotic studies.

(1) Identify the research problem and priorities, design the experiment.

If you are planning to conduct research, you most likely already have a research topic or question in mind. It is important, however, to clarify the research question in such a way as to provide a framework for developing an effective experiment. Typically, this process entails specifying one or more *hypotheses*—predictions about how one factor affects another. The hypothesis for a true experiment can be stated as a prediction that

stipulates that changing the independent variable will cause changes in the dependent variable. For example, you might hypothesize that changing the amount of time subjects practice using an interface will affect how successfully they accomplish tasks with that interface. More specifically, it is important to predict the direction of that change if possible. In this case, we predict that *more* practice will result in *more* successful performance. It is also important to consider how you will manipulate the independent variable. Exactly, what will you change; how will you measure the dependent variable; and what, specifically, counts as better performance? Answering these questions is sometimes called *operationalizing* or developing the *operational definitions* of your variables because you are specifying the operations you will use to manipulate or measure them.

Sometimes, your research question may simply be "I wonder what will happen if…", or "how do people like/use my new interface?" These relatively open-ended research questions are occasionally referred to as fishing expeditions because you do not know what you will catch; they assume that gathering data will provide more insights. They can also be called controlled observation because you would like the people being studied to interact in the same controlled way. This type of question is sometimes criticized for being too general, but for exploratory work, it can be very successful. For example, one of the authors was part of a research team that suspected that subjects confronted with a problem would choose problem strategies based on certain features, but we did not know which features factored into the subjects' choices. So, we included multiple types of problems and multiple types of features. Then, with analysis, we were able to pull out which features subjects relied upon most often based on their decisions across a wide range of problem types (L. M. Reder & F. E. Ritter, 1992).

This book does not describe how to design experiments or how to analyse data. These steps are informed by numerous books on experimental design (some are listed at the end of this chapter). The book also does not explain how to analyse the data, for which again there are numerous excellent books available and noted in further resources.

(2) Develop the experimental task and environment

While it is sometimes interesting to watch what people do when they are left to their own devices, an experiment generally involves giving subjects a specific task: classify these words, solve this problem, answer these questions, and so on. This is even true with controlled observation of a new interface because the test users often have no prior familiarity with the interface. It is important to carefully develop the task you will give your subjects, and to design the environment in which they will perform the task. For example, suppose you want to know how the spacing of learning—whether learning trials are massed into a single session, or distributed over time—affects the ability to retain foreign vocabulary words. To set up this experiment, you would need lists of appropriate words, a means of displaying them, and a way to record the participants' responses. This is typically done with a computer, often programmed using software packages, such as EPrime (Psychology Software Tools, Pittsburgh), specifically designed for behavioral experiments. You would also need to make a large number of practical decisions—how long to display words, how to test memory, and so on. It is especially important to think carefully about how you will collect your data, and how you will verify that your data are being collected correctly. It is very frustrating to spend many hours running an experiment, only to realize that the data were recorded incorrectly (ask us how we know!).

(3) Evaluate potential ethical issues and seek human subjects approval

Once your research protocol is fairly clear, you will want to evaluate your study plan for potential risks or other ethical ambiguities (e.g., whether your experiment examines differences in behavior as a consequence of limited or misinformation). After carefully considering these risks and strategies to mitigate them, you will be ready to seek approval for running human subjects from your organization's IRB or human subjects panel. You should be aware that such approval typically requires extensive paperwork and may take weeks for processing, so you should begin the process as soon as possible and schedule this time in your research agenda. The things you should consider—and that the IRB will want to know—include: how you will recruit and compensate subjects, what you will tell them, whether any personal information will be collected, exactly what they will do doing during your study, what risks subjects might incur by participating, how the data will be kept secure, and so on. Even if you are in a situation in which IRB approval is not required, it is important to think through these issues and plan in advance to address them. We will explain this in more detail in a later chapter.

(4) Pilot test your procedure

A pilot test or pilot study is a preliminary experiment, usually with a small number of subjects, that is intended to let you test the design and procedure of your experiment before you make the investment required to run many subjects. It is an opportunity to make sure that the software running your study works correctly, and that subjects understand the instructions (if there is a way to crash the software or to misinterpret the instructions, rest assured some subject will find it!). You may find that you need to adjust the design or procedure of your study. It is hard to overemphasize the importance of adequate pilot testing. It can be frustrating to take this time when you really want to get onto your "real" experiment, but you will save time and get better results in the long run.

Pilot testing often equates to recruiting whoever is convenient to try out your experiment. For example, you might ask friends and colleagues to try a mirror-tracing task. You might run people casually, in their offices, and record their times to learn how their response times differ by stimuli. You would not report these results, but rather use them to make adjustments. For instance, you may need to alter your apparatus (perhaps adjusting the mirror), your stimuli (you might find how long it takes to follow each pattern), or your procedure (you might have to remind your participants multiple times not to look directly at the sheet of paper).

You will also want to analyze the data you collect during pilot testing. This analysis serves several purposes. First, you will learn whether you have collected data in a format that facilitates analysis, and whether there are any logical gaps between your experiment's design and your plan for analysis. Second, although you won't have a lot of statistical power (because you will have relatively little data), you will get a sense of whether your independent variable "works"—does it actually affect your dependent variable, and whether the relationship is in the direction you hypothesized (or at least in a way that you can interpret)? Finally, you may discover the "edge" of the formal or informal theory guiding your thinking. If, for instance, you are unable to interpret your pilot results using that theory, you may want to consider adjusting the direction of your research.

(5) Prepare an experimental script

During pilot testing, you will likely try various ways of instructing your subjects, as well as some variations to the procedure of your experiment. In your actual experiment, you will want both the manner of instruction and the procedure used to be consistent, to ensure good experimental control. The best way to achieve this is to develop a script for the experimenter(s) running the study to follow. Like a script for a movie or play, this will specify the exact steps to be taken and their sequence. For critical portions, you may want to give your experimenters specific "lines"—instructions that are to be read verbatim to subjects to avoid inadvertently leaving things out or saying them in ways that can be interpreted differently.

(6) Advertise the experiment and recruit subjects

Sometimes, recruiting subjects is easy, and sometimes recruiting is hard. It depends on your local circumstances, the study, and requirements for being a subject. If you have access to a subject pool, it is easier. If, on the other hand, your study requires particular subjects with particular expertise (such as airplane pilots), recruiting is harder.

Recruiting subjects can start while piloting and setting up the study, particularly if preparing the study is relatively easy or recruiting subjects is more difficult. On the other hand, if subjects are easy to recruit and the study is harder to prepare, then recruiting should probably occur after piloting the experiment.

(7) Run the experiment

This is, of course, the heart of the process. Subjects give informed consent, receive instructions, complete your experimental procedure, and are compensated and perhaps debriefed.

Running the experiment may result in different results than those of your pilot study. The primary cause for these differences is generally due to individual variability—participants may think or react in unanticipated ways. Or, you may get different results because your study is more formal. In either of these cases or when there are fewer surprises, you are interested in seeing the truth about the world based on examining a sample of it. How to run the study is the focus of this book.

(8) Analyze the results and archive your study

This is the payoff! If the pilot testing successfully verified the data collection and analysis strategy and the experiment's execution went as planned, you are ready to find out the answer to—or at least better understand—your research question by analyzing the data. The details of data analysis are beyond the scope of this book, but we can offer a few important points that arise while running the experiment.

First, back up your data! If they are on a computer disk, make a copy (or several copies). If they are on paper, photocopy them. Second, make sure the data is stored securely, both to ensure they are retained for use and that they remain confidential. Third, make sure that everything about your experiment is recorded and stored. This task includes archiving a copy of the program that presented materials and collected responses (it's no fun to try to figure out months later which of 5 similarly named files was actually used), a copy of the experimental script, a description of when and where the data were collected, and so on. You may think you'll remember all of these details or that some are not important; but we know from experience that taking the time to carefully archive your

study is worth it because data from an experiment can be used multiple times and much later than they were gathered (e.g., this dataset was analysed multiple ways: Delaney, Reder, Staszewski, & Ritter, 1998; Heathcote, Brown, & Mewhort, 2000; Reder & Ritter, 1988; L. M. Reder & F. E. Ritter, 1992; Ritter, 1989).

(9) Rinse and repeat

Very often—in fact, almost always—your initial experiment is not sufficient to answer your research question. Your data may raise additional questions best addressed by a modification to your procedure, or by evaluating the influence of an additional independent or dependent variable. Though it is not always necessary, conducting additional experiments is often important for understanding your results. Especially if your results are surprising, you may want to repeat the experiment, or part of it, exactly to make sure your results can be replicated.

(10) Report your results

In this step, you take the results and prepare a manuscript. The form used for your manuscript will vary, depending upon your research goals. You may prepare a technical report for a sponsor, a conference paper to test your ideas by exposing them to fellow researchers, a journal article to disseminate novel insights gleaned from your experiment or experiments, or perhaps a thesis to examine a research idea or ideas within a broader context. Regardless of the type of manuscript, you will usually have help and guidance throughout this step (a few resources are noted at the end of this chapter). In addition, there are useful books on the details of the preparation (i.e., *The Publication Manual of American Psychology Association*). While we do not address this topic further in this book, this step is worth keeping in mind throughout the experimental process because reporting your data is not only what you are working towards but also the step that defines many of the requirements associated with the rest of the process (e.g., the emphasis on repeatability).

We have described here an idealized process. The description is normative, in that it specifies what should happen, especially for an inexperienced investigator starting from scratch. In practice, this process often runs in parallel, can vary in order (insights do not always come before or between experiments), and is iterative. Furthermore, breakthroughs frequently result from interactions between researchers across multiple experiments in a lab, so it is usually not a solitary activity.

1.3 Overview of the running examples

We introduce three examples that we will use throughout the course of this book. These examples are hypothetical, but in many cases draw from ongoing or published research. Not all examples will be used in all chapters, but we will use them to illustrate and extend points made in each chapter.

The experimental process begins with the question, "*What do we want to know?"* After determining what we want to know, we must then consider how we go about finding it out. This process entails framing the experiment into either a single or set of falsifiable hypotheses, or, in more complex or emergent situations, simply areas and behaviors where you want to know more. These areas influence the hypotheses, methods, and materials, so it is useful to consider several types of examples.

In the first example, we examine a study from the perspective of a principle investigator working with a special population. We follow Judy, a scientist for an R&D company that cares about

usability. The company specializes in on-demand streaming video. Judy's company is interested in making web-based media more accessible to partially sighted users, users ranging from vision correctable to 20/70 to total blindness. Consequently, she is interested in improving web-navigation for users dependent on screen-readers, a device that provides an audio description of graphical interfaces. To generate this description, screen-readers read the HTML tags associated with the page in question. Consequently, blind users who encounter navigation bars frequently must endure long lists of links. Past solutions have allowed users to skip to the first non-link line; however, Judy is interested in seeing if marking the navigation bar as a navigation feature that can be skipped unless specifically requested improves web navigation for users dependent on screen-readers.

Judy's outputs will include short summaries of results to engineers to more formal technical reports summarizing the whole study and its results. The longer reports may be necessary to describe the context and relatively complex results.

In our second example, we examine a study from the perspective of a graduate student working with an undergraduate. It examines issues in managing less experienced research assistants, and the role of running studies as preparation for writing about them. We will meet Edward and Ying, students at a university. Edward has only recently joined a lab working with e-readers[1] and computer assisted learning tools, while Ying is a Ph.D. candidate in the lab working with the One Laptop per Child (OLPC) project on a small grant. As the OLPC has looked to expand its outreach efforts in the Middle East and Asia, the project has found that common rectangular resolutions produce fuzzy indistinct characters when displaying either Arabic or Hangul. Edward and Ying will be investigating the effects of different screen formats on readability of non-roman alphabets. This research is one component of Ying's thesis examining computer-assisted language learning. Edward will be helping Ying to run subjects in experiments comparing three matrix formats. These formats will differ with respect to pixel density, size, and color. While the experiments are computer-based tests and surveys, Edward will be responsible for greeting participants, explaining the experiments, assisting the participants where necessary and appropriate, and ensuring that the tests have, in fact, successfully recorded the participant's results. To help Edward successfully complete these tasks, Ying will be working with Edward closely.

Ying, Edward, and their faculty advisor will want to continually report and get feedback on their work. How and where they do this informs and can shape the research process, including the details. This process can start with posters at conferences, which provide useful feedback and can help document results. Conference papers (in cognitive science and HCI) provide larger ways to document work, and result in more serious feedback. They will also be interested in Ying's PhD thesis and journal articles on the study.

Our final example focuses on a study done in industry with incremental design as its goal. In this example, we will meet Bob, a human factors engineer, working for a medium-sized company getting into robotics. Bob is simply trying to improve his company's robot in whatever way he can. The platform is both hard to change, while also still flux due to changes being made by the hardware engineers. Bob's situation is the least likely to result in a classic study testing one or more hypotheses. Nevertheless, whether through a study, reading, or controlled observations, he can apply what he learns to improve the human-robot interface (HRI).

In addition, the outputs Bob will be preparing will vary more than the other researchers. Where his engineers are sympathetic and he has learned just a tidbit, he will be able to report what he

[1] This example draws from Al-Harkan and Ramadan (2005).
http://www.sciencedirect.com/science/article/pii/S0169814105000296

learns with a simple e-mail that suggests a change (see, for example, Nielson's comments on types of usability reports used in industry, http://www.useit.com/alertbox/20050425.html). When his engineers are not sympathetic, or where he needs to report more details to suggest larger and more expensive changes, Bob will be preparing reports of usability studies, similar to a technical report. When or if he has general lessons for his field, he may prepare a conference paper or a journal article.

Table 1.2 presents the example studies and where they appear in the book, and what the examples will cover. These studies will be used to provide worked examples and to explore concepts. For example, hypotheses across these studies share some important characteristics: they are usually falsifiable (except in the controlled observation study); and they possess both independent and dependent variables. In examples 1 and 2, there are clear *null* hypothesis: (1) marking the navigation bar to be skipped unless requested *does not help* blind users; and (2) manipulating pixel density, size, and color results *in no difference* in readability. Each hypothesis also has independent and dependent variables. In the first example, a marked or unmarked navigation bar is the independent variable while lag times both within and between the web pages is the dependent variable. For the second example, the independent variables are changes in pixel density, size, and color while the dependent variables are user response times, number of correct responses, and preference rankings. In the third example, the hypotheses are not yet defined. In the third example, Bob would be advised to generate some hypotheses, but his environment may only allow him to learn from controlled observation.

Table 1.2: Summary of example studies used across chapters.

Chapters	**Low Vision HCI Study** *Primary Investigator*	**Multilingual Fonts Study** *Inexperienced RA working with graduate student*	**HRI Study** *Engineer*
Ch. 2 *Preparing the study*	Recruiting/ Special Prep	---	Piloting
Ch. 3 *Ethics*	Stress and Compensation	Ethics and teamwork	---
Ch. 4 *Risks to validity*	---	Internal Validity	External Validity
Ch. 5 *Running the study*	Learning from piloting about apparatus and people	Learning from piloting about people and task	Subject recruitment and when to end the study
Ch. 6 *Concluding a study*	Findings and report	Debriefing Ss, writing up results	Format for reporting, archiving data
Publication Goals	Proof of Concept, Technical Report	Poster Conference Paper PhD thesis	Product changes and future products, Technical Report

Once framed, a study's goals (the testing of a hypothesis, the detection or confirmation of a trend, or the identification of underlying relationships, etc.) inform the rest of the experimental process, all of the steps in Figure 1-1. We provide an overview of each step, as well as discussing these steps in greater detail in the remainder of the book. We will also map this process through our examples. At the end of each section, we will revisit these examples to demonstrate how the information presented in the section translates into practice.

1.4 Further readings

A course in experimental methods is probably the best way to learn about how to design, run, and analyze studies. In addition, we can provide a list of suggested reading materials that provide you with further knowledge about experimental design and methods. We list them in an alphabetical order by first author.

- Bernard, H. R. (2000). *Social research methods: Qualitative and quantitative approaches.* Thousand Oaks, CA: Sage.

  This is a relatively large book. It covers a wide range of methods, some in more depth than others. It includes useful instructions for how to perform the methods.

- Bethel, C. L., & Murphy, R. M. (2010). Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics, 2*, 347–359.

  This article provides practical advice about how to run studies concerning how people use robots. In doing so, it provides a resource that would be useful in many similar studies, e.g., HCI.

- Boice, R. (2000). *Advice for new faculty members: Nihil nimus*. Needham Heights, MA: Allyn & Bacon.

  This book provides guidance on how to teach and research, and importantly how to manage your time and emotions while doing so. Its advice is based on survey data from successful and unsuccessful new faculty members. Some of the lessons also apply to new RAs.

- Coolican, H. (2006). *Introduction to research methods in psychology* (3rd ed.). London, UK: Hodder Arnold.

  If you want a concise text, this book would be a good start. However, it covers all the skills that are required to gently approach research methods.

- Cozby, P. C. (2008). *Methods in behavioral research* (10th ed.). New York, NY: McGraw-Hill.

  Cozby concisely explain methodological approaches in psychology. Also, it provides activities to help you to easily understand the research methods.

- Leary, M. R. (2011). *Introduction to behavioral research methods* (6th ed.). Boston, MA: Pearson.

  As a broad view, this book provides you basic information about a broad range of research approaches including descriptive research, correlational research, experimental research, and quasi-experimental research. It is a comprehensive textbook: you will learn how to proceed through the whole cycle of an experiment, from how to conceptualize your research questions through how to measure your variables to how to analyze the data and disseminate them.

- Martin, D. W. (2008). *Doing psychology experiments* (7th ed.). Belmont, CA: Thomson Wadsworth.

  Martin provides simple “how-to-do” information about experiments in psychology. The author’s informal and friendly tone may help start your journey in this area.

- Ray, W. J. (2009). *Methods: Toward a science of behavior and experience* (9th ed.). Belmont, CA: Wadsworth/Cengage Learning.

  This is a book for the first course in experimental methods in psychology. It is a useful and gentle introduction to how to create and run studies and how to present the results. It does not focus on the practical details like this book does. However, this book can help you learn information about empirically based research and understand cryptic representations in a journal article.

1.5 Questions

Each chapter has several questions that the reader can use as study aids, to preview and review material. There are also several more complex questions that might be used as homework or as topics for class discussion.

Summary questions

1. Describe the following terms frequently used in research with human subjects.
   (a) Stimulus
   (b) Response
   (c) Trials
   (d) Blocks
   (e) Dependent and independent variables
   (f) IRB (Institutional Review Board)
   (g) Statistical power
   (h) Find two more terms and define them.
2. What does "operationalizing the variables" mean in a research study with human subjects?
3. What are the steps in the research process? Create and label a figure showing them.

Thought questions

1. In research with human subjects, having a pilot study is highly recommended. Why do you think a pilot study is important?

2. We have described some database reference tools in university libraries (e.g., PsycInfo, or Web of Science, etc.). Choose any topic you like, and then use these tools to narrow down your research interests or questions. Think about how you would operationalize the variables (i.e., independent and dependent variables) in terms of the topic you just chose. If you find previous studies from the database, compare your operationalized variables with the ones in the published study.

3. Based on the overall research process we described in this chapter, write a summary of the procedures that you need to follow to investigate your topic in the previous question.

4. It is, in general, important to specify operational definitions of the research variables (i.e., independent variables and dependent variables). However, sometimes, it is necessary to gather data to explore a new area of behavior, a so-called fishing expedition. This exploration and data-

gathering can give researchers new, useful insights. As an example, you may look at the Reder and Ritter study in 1992. In this article, Reder and Ritter (1992) designed two experiments testing a range of possible factors influencing feeling of knowing. Discuss what Reder and Ritter observed in the two experiments. Discuss what factors you would explore, how you can manipulate and manage these factors, and how you can measure their effects if you were to run your own study about the feeling of knowing.

5. Within the context of a HCI usability study, discuss what steps in Figure 1-1 you would particularly pay attention to, and which ones you might modify.

2 Preparation for Running Experiments

Often within a lab, multiple experiments are going on at the same time. Joining the lab as a new research assistant, you have come to help out and to learn in this area, specifically with running research studies. What do you do? Where do you start? How do you avoid common and easily fixed problems? This chapter describes how to get started. Figure 2-1.

Consider briefly a usability study evaluating a haptic (touch-based input or output) interface. For this investigation, a lead research scientist or a lead researcher would establish a study hypothesis and design an experiment by first defining what to measure (dependent variables), what factors to manipulate (independent variables), and what environmental conditions to consider. This work would be piloted and would take some time to prepare.

The whole preparation process is represented in Figure 2-1, along with the section (§) or sections (§§) that explain that step.

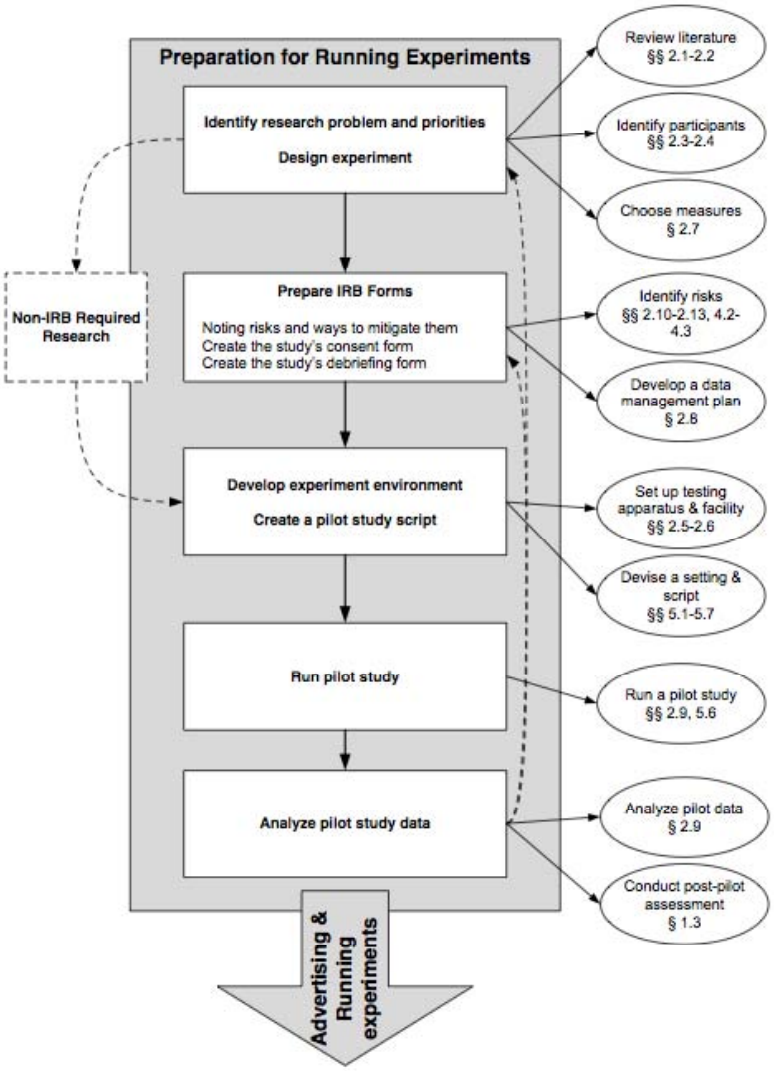


Figure 2-1. A pictorial summary of the study preparation process.

2.1 Literature in the area

This book does not assume that you have a background in statistics or have studied experimental design. To help run a study, you often do not need to be familiar with these topics (but they do help!). If you need help in these areas, there are other materials that will prepare you to design experiments and analyze experimental data, which are noted at the end of this chapter. In addition, most graduate programs with concentrations in HCI, cognitive science, or human factors feature coursework that will help you become proficient in these topics.

Many introductory courses in statistics, however, focus primarily on introducing the basics of ANOVA and regression. These tools are unsuitable for many studies analyzing human subject data where the data is qualitative or sequential. Care, therefore, must be taken to design an experiment that collects the proper kinds of data. If ANOVA and regression are the only tools at your disposal, we recommend that you find a course focusing on the design of experiments featuring human participants, and the analysis of human data. We also recommend that you gather data that can be used in a regression because it can be used to make stronger predictions, not just that a factor influences a measure, but in what direction (!) and by how much.

Returning to the topic of readings, it is generally useful to have read in the area in which you are running experiments. This reading will provide you further context for your work, including discussions about methods, types of subjects, and pitfalls you may encounter. For example, the authors of one of our favorite studies, an analysis of animal movements, notes an important pitfall, that data collection had to be suspended after having been chased by elephants! If there are elephants in your domain, it is useful to know about them. There are, of course, less dramatic problems such as common mistakes subjects make, correlations in stimuli, self-selection biases in a subject population, power outages, printing problems, or fewer participants than expected. While there are reasons to be blind to the hypothesis being tested by the experiment (that is, you do not know what treatment or group the subject is in that you are interacting with, so that you do not implicitly or inadvertently coach the subjects to perform in the expected way), if there are elephants, good experimenters know about them, and prepared research assistants particularly want to know about them!

As a result, the reading list for any particular experiment is very individualized. You should talk to other experimenters, as well as the lead researcher about what you should read as preparation for running or helping run a study.

2.2 Choice of a term: Participants or subjects

Disciplines vary as to which term they prefer: *subject* or *participant* and how the role of the people you study is not completely passive. *Participant* is the newer term, and was adopted by many research communities to emphasize the researcher's ethical obligations to those participating in their experiment. Even more descriptive terms such as learner, student, or user can be used and are generally preferred. Nevertheless, *subject* is still commonly used, and appears in older research. For students in many psychology programs, the term, *participants,* is preferred by some to that of *subjects*. The *Publication Manual of the American Psychological Association (APA)*, 5th ed. (American Psychological Association, 2001, p. 70) suggests replacing the impersonal term, *subjects*, with the more descriptive term, *participants*. The APA goes on to define *participants* as individuals: college students, children, or respondents. The APA manual suggests this, but does not require it.

Indeed, the *Publication Manual of the APA* (6th ed.) stops far from requiring the use of 'participants'. It says this about the use of the term "subjects":

> Write about the people in your study in a way that acknowledges their participation but is also consistent with the traditions of the field in which you are working. Thus, although descriptive terms such as college students, children, or respondents provide precise information about the individuals taking part in a research project, the more general terms participants and subjects are also in common usage. Indeed, for more than 100 years the term subjects has been used within experimental psychology as a general starting point for describing a sample, and its use is appropriate. (p. 73)

No matter how you write with respect to the APA guidelines, we should recognize that *S*, *Ss*, *S's*, *E*, *Es, E's* indicate *Subject*, *Subjects, Subject's*, *Experimenter*, *Experimenters,* and *Experimenter's* in earlier research—Fitts's 1954 study is one example where these abbreviations are used. Furthermore even within the discipline of psychology, opinion can be split. Roediger (2004) argues against the change to *participants* suggested in the latest version of the *APA's Publication Manual*. He argues that *subjects* is both more consistent and clearer, noting that the term has been in use since the 1800's and that it better defines the relationships involved. He argues that the term, *participants,* fails to adequately capture the distinction between the experimenter and those in the study—strictly speaking experimenters are participants as well.

We use these terms interchangeably in this document because we recognize and respect the fact that other research communities may still prefer *subjects*, and because not all psychologists, and certainly not everyone running behavioral experiments, are members of the American Psychological Association.

Another distinction[2] to draw in this area is what the purpose of the study is. If the topic of interest to you is a psychological phenomenon, an aspect of human behavior, the people in your study may appear more as subjects in the traditional use of the term. On the other hand, it may be that you are actually interested in how someone performs when given a certain interface or new tool and task. In this case, you are actually interested in how well the widget works. Consequently, your subjects are really more like participants who are participating with you in your work, helping you to generalize results and to improve the product. In any case, take advice about what to call the people you work with.

2.3 Recruiting participants

Recruiting participants for your experiment can be a time consuming and potentially difficult task, but it is a very important procedure to produce meaningful data. An experimenter, thus, should carefully plan out with the lead researcher (or the principal investigator) to conduct successful recruitment for the study. Ask yourself, "What are the important characteristics that my participants need to have?" Your choices will be under scrutiny, so having a coherent reason for which participants are allowed or disallowed into your study is important.

First, it is necessary to decide a population of interest from which you would recruit participants. For example, if an experimenter wants to measure the learning effect of foreign language vocabulary, it is necessary to exclude participants who have prior knowledge of that language. On the other hand, if you are studying bilingualism you will need to recruit people who speak two languages. In addition, it may be necessary to consider age, educational background, gender, etc., to correctly choose the target population.

Second, it is necessary to decide how many participants you will recruit. The number of participants can affect the ability to generalize from your final results. The more participants you

[2] We thank Karen Feigh for this suggested view.

can recruit, the more reliable your results will be. However, limited resources (e.g., time, money, etc.) force an experimenter to find the appropriate and reasonable number of participants. You may need to refer to previous studies to get some idea of the number of participants, or you may need to calculate the power of the sample size for the research study, if possible (most modern statistical books have a discussion on this, and teach you how to do this, e.g., Howell, 2008). Finally, you will upon occasion have to consider how many are too many. Running large numbers of subjects can waste both time and effort. In addition, the types of statistics that are typically used become less useful with larger sample sizes. With large sample sizes, effects that are either trivial or meaningless in a theoretical sense become significant (reliable) in a statistical sense. This is not a normal problem; but if, for example, you arrange to test everyone in a large class you could potentially encounter this problem.

There are several ways that participants can be recruited. The simplest way is to use the experimenters themselves. In simple vision studies, this is often done because the performance differences between people in these types of tasks is negligible, and knowing the hypothesis to be tested does not influence performance. Thus, the results remain generalizable even with a small number of participants.

Subjects can also be recruited using samples of convenience. Samples of convenience consist of people who are accessible to the researcher. Many studies use this approach, so much so that this is not often mentioned. Generally for these studies, only the sampling size and some salient characteristics are noted that might possibly influence the participants' performance on the task. These factors might include age, major, sex, education level, and factors related to the study, such as nicotine use in a smoking study, or number of math courses in a tutoring study. There are often restrictions on how to recruit appropriately, so stay in touch with your advisor and/or IRB.

In studies using samples of convenience, try distributing an invitation email to a group mailing list (e.g., students in the psychology department or an engineering department) done with approval of the list manager and your advisor. Also, you can post recruitment flyers in a student board, or an advertisement in a student newspaper. Use efficiently all the resources and channels that are available to you.

There are disadvantages to using a sample of convenience. Perhaps the largest is that the resulting sample is less likely to lead to generalizable results. The subjects you recruit are less likely to represent a sample from a larger population. Students who are subjects are different from students who are not subjects. To name just one feature, they are more likely to take a psychology class and end up in a subject pool. And, the sample itself might have hidden variability in it. The subjects you recruit from one method (an email to them) or from another method (poster) may be different. We also know that they differ over time — those that come early to fulfill a course requirement are more conscientious than those that come late. So, for sure, randomly assign these types of subjects to the conditions in your study.

The largest and most carefully organized sampling group is a random sample. In this case, researchers randomly sample a given population by carefully applying sampling methodologies meant to ensure statistical validity and equal likelihood of selecting each potential subject. Asking students questions at a football game as they go in does not constitute a random sample—some students do not go (thus, creating a selection bias for those subjects who like football, who have time and money, and certain interests, etc.). Other methods such as selecting every 10^{th} student based on a telephone number or ID introduce their own biases. For example, some students do not have a publicly available phone number, and some subpopulations register early to get their ID numbers. Truly choosing a random sample is difficult, and you should discuss how best to do this with your lead researcher.

In any case, you need to consider what subjects you will recruit and how you will recruit them because you will need to fill in these details when you submit your IRB forms (covered later in this chapter).

2.4 Subject pools and class-based participation[3]

One approach for recruiting participants is a *subject pool*. Subject pools are generally groups of undergraduates who are interested in learning about psychology through participation in experiments. Most psychology departments organize and sponsor subject pools[4].

Subject pools offer a potential source of participants. You should discuss this as an option with your lead researcher, and where appropriate, learn how to fill out the required forms. If the students in the study are participating for credit, you need to be particularly careful to record which students participated and what class they are associate with because their participation and the proof of that participation represent part of their grade.

A whole book could be written about subject pools. Subject pools are arrangements that psychology or other departments provide to assist researchers and students. The department sets up a way for experimenters to recruit subjects for studies. Students taking particular classes are either provided credit towards the class requirement or extra credit. When students do not wish to participate in a study, alternative approaches for obtaining course credit are provided (but are rarely used).

The theory is that participating in a study provides additional knowledge about how studies are run, and provides the participant with additional knowledge about a particular study. The researchers, in turn, receive access to a pool of potential subjects.

Sometimes, researchers can make arrangements with individual instructors to allow research participation for extra credit in a course (drawings or free food do not seem to encourage participation). In such cases, it is important to keep in mind some of the lessons learned by those who have run subject pools. These lessons will be important in receiving IRB approval[5] for this approach to recruiting. First, the amount of extra credit should not be too large, both to avoid coercion to participate (by offering too great an incentive) and to avoid compromising the grading scheme for the class. Second, an alternative means of earning extra credit must be provided. The alternative assignment should be comparable in time and effort to participating in the study. For example, students might read a journal article and write a 2 or 3 page summary rather than participating in the study. Usually the researcher, rather than the instructor, must take responsibility for evaluating the alternative opportunity to keep instruction separate from the participation in research, but this can raise separate issues.

As in using a subject pool, it is important to have a research protocol that provides a secure record of research participation and a procedure for sharing that record with instructors that ensures students receive credit while maintaining confidentiality. For example, it is often best to ask for student ID numbers in addition to names on consent forms to avoid difficulties due to illegible handwriting. If you are providing extra-credit opportunities for students in multiple classes, you should plan for the possibility that some students are taking more than one class, so that you can avoid having the same person participate in your study twice. Multiple participation poses

[3] Some of the ideas in this section are taken from an email from in the College of IST, November, 2010.
[4] Note, the APA does not appear to call for calling these participant pools.
[5] Further information is available from your IRB, for example, http://www.research.psu.edu/policies/research-protections/irb/irb-guideline-5

problems both of ethics (receiving double credit) and internal validity (non-independent observations). It is also appropriate to end the study about two weeks before the end of the semester to allow time to enter grades and resolve inconsistencies.

Subject pools usually have written rules and procedures designed to help researchers with these issues. It is important to learn these rules and procedures, and to follow them carefully, to avoid possible problems.

2.5 Care, control, use, and maintenance of apparatus

What materials do you need to run experiments? The experiments in a controlled environment (e.g., a laboratory) usually require participants to interact with a computer device, a prototype, or a mock-up. For example, it is possible to implement a task environment in a computer screen—such as an air traffic control task like Argus (Schoelles & Gray, 2001), a driving simulator like Distract-R (Salvucci, 2009), experimental tasks with E-Prime (e.g., MacWhinney, St. James, Schunn, Li, & Schneider, 2001), or a spreadsheet task environment (J. W. Kim, Koubek, & Ritter, 2007).

Part of what you will have to do to set up and run a study is to understand the task environment so that you can prepare it for each session, save the data if it collects data, and shut it down after each session.

As you begin to work on your research task, you are likely to consider several approaches for improving your study. Finding, developing, or modifying the task environment to support your study is often an early consideration. The task environment provides the setting for investigating the questions of interest, and having the right task environment is a key element to a successful study. If designing and implementing a new task environment for your research study seems infeasible, try reusable and sharable environments. With the increasing use of computerized task environments, this is increasingly possible. For example, Argus is available (Schoelles & Gray, 2000), and there are multiple versions of games such as SpaceFortress (Mané & Donchin, 1989; Moon, Bothell, & Anderson, 2011) and other games on the Internet.

After choosing and setting up the task environment, the next step is to determine what method you will use to record the participant's performance. Data collection deserves serious thought. Data can be qualitative (i.e., not in a numerical form) or quantitative (i.e., in a numerical form). Different hypothesis and theories require different types of data to test them, and thus methods to collect data. For example, you can use a camcorder in an interview to gather qualitative information or a keystroke logger like RUI (Kukreja, Stevenson, & Ritter, 2006) to measure numerical values of quantitative data in unobtrusive and automatic ways. We suggest avoiding manually recording data—it is hard, takes a significant amount of time, and is prone to error. Though, sometimes, manual data collection is unavoidable and for pilot studies it is quite often appropriate. Often with a little forethought ways can be found to automate the process.

An apparatus is often required to gather behavioral data. In cognitive science, recording user behavior by using experimental software, a video recorder, a voice recorder, or a keystroke/mouse logger, etc are all common practices. There are also tools for generating studies such as ePrime. Also, some studies require using an eyetracker to gather eye movement data.

2.5.1 Experimental software

Many studies are performed with custom built, or proprietary software. The research team conducting the study usually develops these custom applications. They can vary from a simple program to present stimuli and record reaction times to more complex programs (interactive simulations for instance). As a new research assistant, you will be instructed on how to start up

and run the software necessary for your work. On the other hand, as you run subjects with such programs, try moving from a passive to an active user. Make suggestions that you think might improve the program's usability as they arise, note mistakes in the program, and observe how subjects interact with the program in novel or interesting ways. These insights can lead to further studies and to further hypotheses to test.

2.5.2 E-Prime

E-Prime[6] was the first commercial tool designed to generate psychological experiments on a personal computer (MacWhinney, St. James, Schunn, Li, & Schneider, 2001). E-Prime is compatible with Microsoft Windows® XP/Vista. PsyScope[7] is another experiment generation program, and a predecessor of E-Prime. You can download PsyScope free under a GNU General Public License[8]. PsyScope runs on the Macintosh. You may be asked to use these tools in your current study or may find them to be of great value in producing study stimuli more quickly.

2.5.3 Keystroke loggers

It is often useful to record the user's behavior while they perform the task, not just the total task time. This can be done in several ways. Some researchers have used video recordings. This provides a very stable result that can include multiple details. It also can provide a rich context, particularly if both the subject and their surroundings are recorded. On the other hand, analyzing video recordings is time consuming and can be error prone. Analyzing video data often requires examining the video frame-by-frame to find when the user performs each action, and then recording each action by hand into your dataset.

Another approach is to record just the keystrokes or mouse clicks. There are commercial versions available from companies like Noldus that will record keystrokes. We have also designed a keystroke logger, RUI (Recording User Input). RUI is a keystroke and mouse action logger for the Windows and Mac OS X platforms (Kukreja, Stevenson, & Ritter, 2006). It is a useful tool for recording user behavior in human-computer interaction studies. RUI can be used to measure response times of participants interacting with a computer interface over time.

Figure 2-2 shows an example output from RUI. It includes a header to the file noting who the subject was, and the date of the log. There is a header line noting the column contents, with time in elapsed time rather than HH:MM:SS.mmm (the elapsed time seems to work better). You might create similar logs if you instrument your own system.

Using RUI or other keystroke loggers, however, can raise issues regarding privacy in public clusters (e.g., a classroom). University policies almost universally prohibit installing any tool for experimentation that obtains or could obtain a user's information on identity such as a login ID or a password (J. W. Kim & Ritter, 2007). Fortunately, Kim and Ritter (2007) describe one possible portable solution to this problem. They used a simple shell script to automatically run RUI on an external drive, a jump drive. When RUI is operated from an external drive it provides a way to efficiently use RUI on public cluster machines and then remove it when the study is over. A later version of RUI anonymises the keystroke values.

[6] http://www.pstnet.com/products/e-prime
[7] http://psy.ck.sissa.it
[8] http://www.gnu.org/copyleft/gpl.html

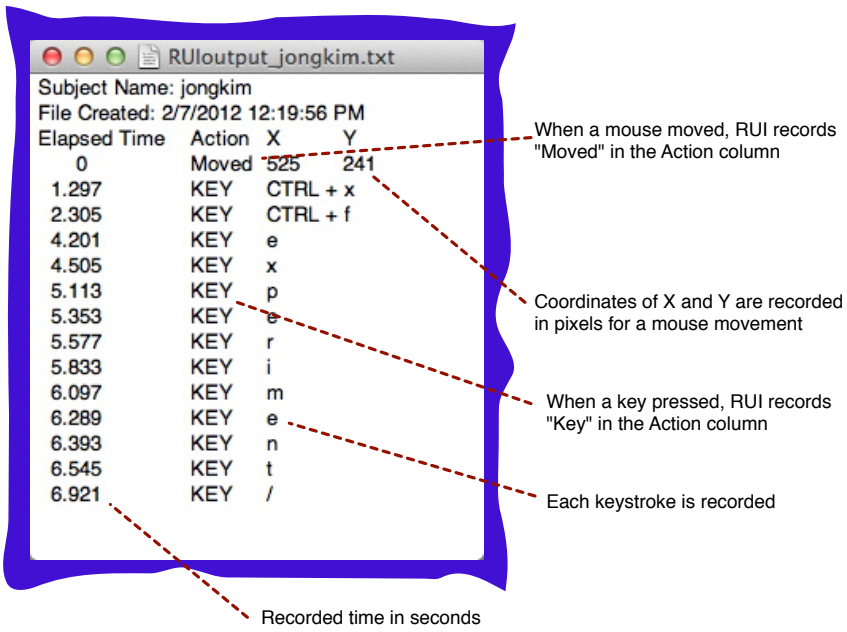


Figure 2-2. A screenshot of logged data recorded in RUI.

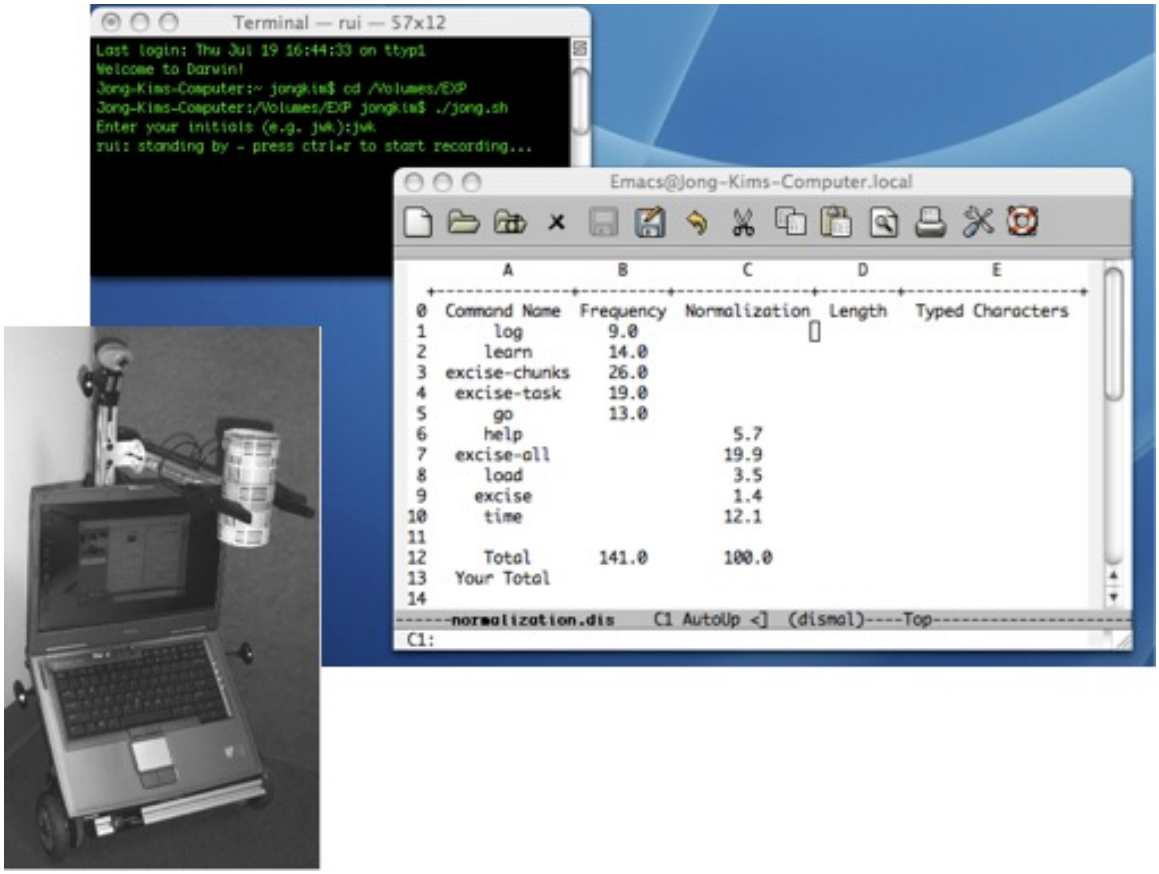


Figure 2-3. Interfaces that RUI can be run on (ER1 robot and the Dismal spreadsheet)

2.5.4 Eyetrackers

An eyetracker is a device to record eye positions and movements. In general, a researcher generally analyzes the recorded eye movements that are a combination of two behaviors: (a) fixations—pauses over informative regions that are of interest, and (b) saccades—rapid movements between fixations (Salvucci & Goldberg, 2000). It can offer useful data about the cognitive processes (Anderson, Bothell, & Douglass, 2004; e.g., Salvucci, 2001) when a user interacts with an interface (e.g., a computer screen, a physical product, etc). This apparatus is sensitive, requiring special care to guarantee the measurement's quality, but they are becoming easier to use and less expensive over time.

Figure 2-4 shows someone wearing a head-mounted eye-tracker. To the right of the computer display are three monitors showing how well the eye is being tracked, what the scene camera is viewing, and the scene camera with the eye's position superimposed. The bar on the right is used to track a metal plate in the hat, and thus track where the head and eyes are pointed.

Figure 2-4. Subject wearing a head-mounted eye-tracker. (Photo by Ritter.)

2.6 The testing facility

A testing facility can be called a psychological testing room, human factors lab, an ergonomics lab, a usability lab, or a HCI lab. Rosson and Carroll (2002) describe a usability lab as a specially constructed observation room. In this observation room, an investigator can simulate a task environment and record the behavior of users. Thus, the room should be insulated from outside influences, particularly noise. However, it is sometimes necessary to observe and record behaviors of a group of users interacting with each other. In these cases, it may be hard to capture this data in a lab setting. Ideally, the testing facility should be flexible enough to conduct various types of research.

Jacob Nielson (1994) edited a special journal issue about usability laboratories. This special issue provides several representative usability laboratories in computer, telecommunications, and consumer product companies (e.g., IBM, Symantec, SAP, Phillips, or Microsoft, etc.). While this special issue is somewhat dated, the underlying concerns and some of the technological details

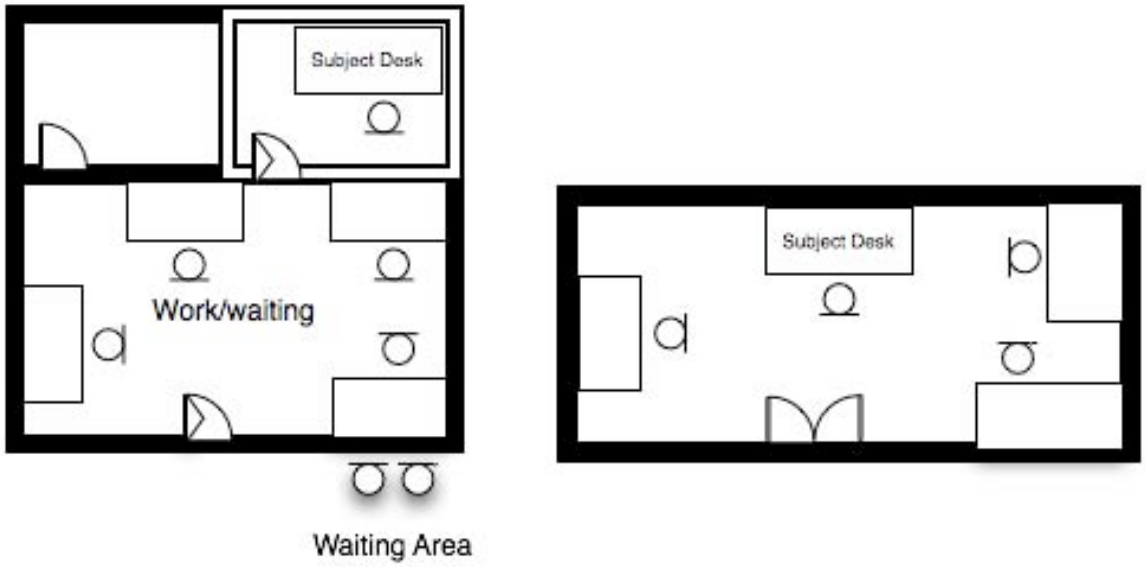
remain accurate; in addition, many of the social processes and uses for video have only become more important.

If you are designing your own study, you should try to arrange access to a room that allows participants to focus on the experimental task. Lead researchers will often have such rooms, or can arrange access to them.

Figure 2-5 shows two different spaces. The space on the left has a room that was built to provide sound isolation by including fiberglass insulation between two sheets of particleboard. The doors into the lab and into the running room have sweeps on them to further keep out noise. The entry room might be a room where students work, but it provides some quiet and a natural place to welcome and debrief the subjects. There are chairs for subjects to wait at outside the room if they are early, and (not shown) its room number is clearly shown by the door.

The right hand side of Figure 2-5 is a poor room to run studies in. The subject is in a room where people will be working, and thus they can get distracted while doing their task. There is no place to wait, and because their back is to two doors whenever someone comes in from the hallway they will be tempted to turn and look at them, causing noise in the data.

We offer further advice on the setup of your experimental space in Chapter 5, on running an experimental study.



Lab space that supports running studies Lab space with less support for running studies

Figure 2-5: Example diagrams of space for running studies.

2.7 Choice of dependent measures: Performance, time, actions, errors, verbal protocol analysis, and other measures

The point of conducting an experiment is to observe your subjects' behavior under controlled conditions. Prior to beginning your experiment, it is important to consider exactly what it is you want to observe, and how you will measure it so that you can effectively summarize your

observations and conduct statistical tests. That is, you must choose your dependent variables and decide how to measure them.

2.7.1 Types of dependent measures

A very common kind of observation is simply whether or not the subject succeeds at performing the task. Often, this is a yes-or-no question, and you might summarize your data by calculating the proportion of your subjects who succeed in different conditions (that is, at different levels of your independent variable). If the task requires repeated responses from each subject, you might calculate the proportion (or percent) of correct responses for each subject. For example, if the focus of your study is memory, your measure might be the proportion of items correctly recalled or recognized. It is important to think carefully about the measure you use. In the case of memory, for instance, you may find dramatically different results depending on whether you measure recognition or recall. Not only is recognition generally easier than recall, some independent variables will have different effects depending on which measure of memory you choose. Furthermore, if you choose to measure recall, the type of cue you provide to prompt subjects to recall will make a difference in your results.

Sometimes, determining whether or not subjects succeed at your experimental task requires a judgment call. For example, suppose you are interested in whether subjects successfully recall the gist of instructions presented in each of several interfaces. While it would be simple to calculate the proportion of exact words recalled, that would fail to capture the observation of interest. In such cases, you need to make an informed judgment about the success of recall. In fact, you should have two or more people make such judgments, to determine the reliability of the judgments. Your judges should make their decisions "blind"—that is, they should not know which experimental condition a subject was in, so that they cannot be unwittingly influenced by their knowledge of the hypothesis.

In many cases, experimental tasks are designed so that almost every subject succeeds—responds correctly—on almost every trial. In such cases, the time to respond, often known as *reaction times or response times*, can provide a more sensitive measure of performance. For almost all tasks, faster is better, as long as performance is accurate. There are exceptions, of course—the pianist who plays a song the fastest may not be the one who best succeeds at conveying the musical message. When using response time measures, it is also important to consider the possibility of a *speed-accuracy tradeoff*—subjects may achieve faster performance by sacrificing accuracy, or vice versa. Usually, it is easiest to interpret response time if the conditions that lead to faster performance also lead to greater accuracy. And sometimes, how subjects choose a speed-accuracy tradeoff may be of great interest.

Another kind of dependent measure is a self-report. Questionnaires are one common and flexible way to collect self-reports. By answering the questions, participants self-report about the question, thus providing researchers insights into their behavior. The quality and type of these responses, however, depend upon the quality and type of the questions asked—so carefully selected and carefully worded questions are important. One example where questionnaires can be used effectively is studying self-judgment and its effects. Under certain conditions, our feelings about our knowledge and our actual knowledge may differ. In this case, our hypothetical researcher may ask the participants to make a judgment about what they know after memorizing vocabulary words. To measure the participants' self-judgment, he or she could use a Likert scale. Likert scales are one common approach, and typically consist of five to seven points with ratings ranging from "Strongly disagree" to "Strongly agree". Our hypothetical researcher would then test the participants and compare the participants' responses about their knowledge with the results.

Another type of data to gather is error data. Error data consists of trials or examples where subjects did not perform the experimental task or some aspects of the task correctly. This type of data can provide useful examples of where cognition breaks down. In addition, it helps describe the limits of performance and cognition.

Error data is generally more expensive to collect because in most cases participants perform the task correctly. Thus, generally more trials have to be run to gather a hundred errors than it takes to gather a hundred correct responses. Conversely, if errors are not of interest to your research topic, some pilot running of the experiments may be required to generate an experiment where errors do not occur too often.

The measures we have discussed so far all reflect the outcome of behavior—the final result of a psychological process in terms of successful performance, time, or subjective experience. Often, however, research questions are best addressed by *protocols* or *process tracing* measures—measures that provide insight into the step-by-step progression of a psychological process. One example is recording the sequence of actions—moves in a problem solving process, the timing and location of mouse clicks when using a computer interface, and so on. Computerized task environments make it relatively easy to collect such measures, though aggregating and interpreting them can be challenging. Figure 2-6 shows a trace of where people finding a fault in a circuit look at the interface.

Sometimes tasks are designed especially to allow the interpretation of action sequences in terms of theoretical questions. For example, Payne and his colleagues (e.g., Payne, Braunstein, & Carroll, 1978) recorded the sequence of information-acquisition actions to generate evidence about decision strategies and processes. Protocols may include multiple streams of data including verbal utterances, motor actions, environmental responses, or eye movements (Newell & Simon, 1972). As an example of a verbal protocol, consult the testing methodology developed by Ritter and Larkin (1994) for the principled analysis of user behavior. Protocol data cannot be reduced to simple averages, but can be used in a variety interesting ways to provide insight into behavioral processes **(**Sanderson & Fisher, 1994). Often, protocol data are analyzed by comparing them to the predictions generated by computational models intended to simulate the process being studied.

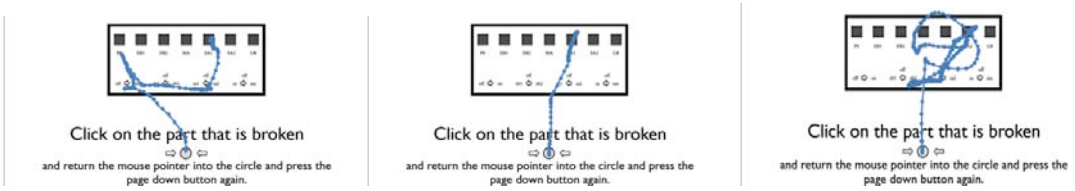


Figure 2-6. Example eye-tracking traces of problem solving showing rather different strategies solving the same problem. (taken with permission from Friedrich, 2008.)

Verbal protocols often provide insights into understanding human behavior. Ericsson and Simon (1993) published a summary of how and when to use verbal reports as data to observe humans' internal cognitive processes. The basic assumption of their verbal protocol theory is that verbalization of a human's memory contents (not their view of their thought processes) can be used to derive the sequence of thoughts to complete a task. The basic distinction they make is between *talking aloud*, in which subjects simply say what is in mind as they perform a task, and *thinking aloud*, which involves reflection on mental processes. Talking aloud data is generally more valid because it is less likely to be contaminated by the subject's theories of his or her own behavior. Thus, verbalization can be a valid form of data that offers unique insights into

cognition. For example, in a learning experiment, subjects can often report the hypotheses they are considering, but reporting *why* they are considering a particular hypothesis is likely to depend on their naïve theories of behavior and is much less valid. It is also important to consider whether the processes being studied are actually represented verbally—much of our thinking is in a verbal format and thus is easy to report, but a task that is carried out primarily on the basis of visual imagery is not suitable for verbal protocols. Not only is the need for the subject to translate his or her visual images into words a source of error, verbalizing such tasks often interferes with performing the task (e.g., Schooler, Ohlsson, & Brooks, 1993). Work in this area has, for example, helped us understand how experts play chess (de Groot & Gobet, 1996).

Collecting verbal protocol data requires audio recordings, and often comes with special apparatus for recoding and special software and tools for analyzing the results. Collecting, transcribing, and coding such data is very time consuming, but can be very helpful for understanding how the task is performed. It is especially important to have a clear plan for analyzing protocol data, and to link the data to the actual behavior observed in the experimental task. In the 1980s and 1990s, the newly-respectable use of verbal protocols provided great insight into complex tasks such as problem solving. These successes encouraged many researchers to collect verbal protocols, often without sufficient forethought. One of us has several times had the experience of trying—with very limited success—to help a researcher who collected large amounts of verbal protocol data without any plan for analyzing it. In one case, many hours of data collection and transcription were useless because there was no way to link the verbal reports to the actual behavior!

Physiological measures can also provide insight into behavior, though they require substantial investments in equipment and technical skills. Cozby (2004) introduces a few popular physiological measures such as galvanic skin response (GSR), electromyogram (EMG), and electroencephalogram (EEG) that help us understand psychological variables. Also, fMRI (functional magnetic resonance imaging) is a popular method of measuring and examining brain activities. If you are interested in learning more about these techniques, refer to the section of Further Readings, specifically *Psychophysiological Recording* (Stern, Ray, & Quigley, 2001).

2.7.2 Levels of measurement

Often within a single study, multiple measures with different characteristics are gathered. Let us discuss some common measures taken in an HCI or cognitive science experiment. For instance, you can measure the task completion time; or you can measure the number of keystrokes and mouse actions performed by the participants during the task, as well as the timestamp associated with each action. You can also measure what errors were made during the task, and so on.

It is necessary to decide what you are observing and measuring from the participants who are performing the experimental task. The decision is important because the choice of measures is directly related to what aspects of the participants' behavior you are trying to capture in the task. In general there are two types of variables: (a) independent variables, and (b) dependent variables.

Independent variables cause, or manipulate the changes in the participants' behavior that the researchers seek to observe during the study. Thus, independent variables are sometimes called manipulated variables, treatment variables, or factors (Keppel & Wickens, 2004).

To cement our understanding of variables, let us presume that we want to measure how humans forget something they have learned. We will return to this example later, but for now, we will focus on the study's independent and dependent variables. Variables that can manipulate forgetting performance include training types, retention intervals (how long a participant will retain learned information), and input modalities (what types of skills a participant is to learn). Thus, we would consider these variables the study's independent variables. They are deliberately varied to create the effects—they are independent. Variables that are fixed going in, such as

gender, sex, age, are also treated as independent variables because they are not dependent on the treatment.

Dependent variables indicate what we will observe. Their values are (presumed to be) dependent on the situation set up by the independent variables. Dependent variables can either be directly observed or may be derived. Response time and error rates are two typical dependant variables. The measures can also be more complex. Workload measures for example, allow researchers to measure how hard users have to work. The NASA TLX (Hart & Staveland, 1988; NASA, 1987) directly measures workload using six individual subscales, but sometimes a desired measure is used based on combining them. We can observe the time that is required to complete a task if the investigation is to understand human performance caused by forgetting. Also, we can observe errors produced by participants to measure forgetting. These variables are considered to be dependent variables. There can be one or more dependent variables. One dependent variable in an experiment uses univariate statistical methods, and more than two dependent variables require multivariate methods.

To sum up, dependent variables are the responses being observed during the study while independent variables are those factors that researchers manipulate to either cause or change those responses.

2.7.3 Scales of measurement

Variables can be measured using four types of scales (Ray, 2003): (a) nominal measurements, (b) ordinal measurements, (c) interval measurements, and (d) ratio measurements. Knowing these scales of measurement is important because the data interpretation techniques available to you for interpreting the results are a function of the scales of measurement used, and the use of such data, perhaps even how it is stored and the way equipment is calibrated can depend on what kind of data it is.

Nominal (also referred to as categorical) measurements are used to classify or name variables. There is no numeric measure of values representing names or separate categories. For example, participants can be classified into two groups—a male group and a female group, to measure performance on using a GPS navigation system. In this case, the gender difference is an independent categorical variable to compare performance. Or, if the numbers 1 to 10 are treated as words, such as how often they are said, then there is not necessarily even an order to them, they could be sorted alphabetically.

Ordinal measurements, in contrast, represent some degree of quantitative difference (or relative amount). For example, football rankings in the Big Ten conference are an ordinal measurement; they are in order, as are ratings on a scale of 1 to 10. Differences between the first and second team, between 9^{th} and 10^{th}, and between ratings of 4 and 5 and 6 and 7 are not necessarily equal, just ordered.

Interval measurements rely upon a scale values based on a single underlying quantitative dimension. The distance, therefore, between the consecutive scale values are meaningful. For example, the interval between 6 and 12 is the same as the interval between 12 and 18.

Ratio measurements determine values with respect to an absolute zero—there is no length shorter than 0 inches for instance. The most common ratio measurement can be found in a count measures (i.e., the number of hits or misses). For example, in a shooting game, the number of hits is used to determine the shooter's accuracy.

It is important to understand the scales of measurement of your variables for several reasons. First, the scale of measurement determines the mathematical operations you can perform on your data. For example, if you code male subjects as 0 and female subjects as 1, it makes no sense to

say that the average gender was 0.3; instead, you would report the actual numbers or proportions. Similarly, while averaging ordinal data seems to make sense, because the intervals may not be equal, an average ranking of 4 is not necessarily twice an average ranking of two. Second, as a consequence of these limits on mathematical operations, different statistical techniques are required for data on different scales of measurement. Parametric statistics, which include such common tests as analysis of variance, require at least interval measurements. Ordinal or nominal scale data should be analyzed using *non-parametric statistics* such as the chi-square (χ^2) test.

2.8 Plan data collection with analysis in mind

It is quite easy to record data using computer software, audio and video recording equipment, or even pencil and paper. Recording data in a way that makes it easy to analyze can be a bit more challenging. You will save a great deal of time and effort, and perhaps avoid the need to repeat the experiment, if you keep these points in mind when planning your data collection:

- Record everything you will need, including the appropriate identifiers for your data. It is important to capture everything you will want to know about each subject's participation, and doing so requires some thought. For example, you may want to record the time of each key press by the subject; in this case, make sure that you know exactly which event begins the timing that is recorded. If you collect some of your data using a computer and some by paper-and-pencil, make sure that you have a foolproof method of matching the computer data file with the appropriate paper-and-pencil data. If your stimuli are randomized, make sure that you record which stimulus the subject saw on each trial, so that it can be matched with the appropriate response. It may seem obvious what should be recorded, but our experience suggests that it is important to think this through carefully. You will never regret recording some aspect of the data—if it turns out to be irrelevant, you don't need to analyze it—but it is impossible to go back and recover observations you didn't make. We know from experience that experiments sometimes have to be repeated because some part of the data that turned out to be critical was not recorded.
- Organize the data appropriately for analysis. Data analysis software generally expects data to be in a format similar to that of a spreadsheet, in which each line of data in the file represents one case. Make sure that each line of data includes the appropriate identifiers—subject number, level of each independent variable, the specific stimulus displayed if relevant, and so on. Plan your data file so that each line of data includes the same number of values. If different trials have different numbers of variables—for example, in an experiment on working memory where different trials may require subjects to remember different numbers of items—plan codes to indicate that some variables are not relevant to some trials. One of us recently neglected this, and consequently had to spend many hours reorganizing the data from a study! Plan data entry carefully.
- Choose an appropriate format for data storage. The chances are good that you will find yourself transferring data from one program to another (for example, from EPrime to SPSS, or from Excel to SPSS). A very common format for data storage is Comma Separate Values (CSV), in which each line of the data file consists of a list of numbers separated by commas. Most spreadsheet and statistical programs can easily read this format. Most programs will also accept similar formats in which spaces, tabs, or sometimes other characters separate values instead of a comma.
- Plan for data entry. Often, some of your data, perhaps all, will have to be entered into your data file by hand. This task, while mundane, is error prone. Make sure that such data is collected in a way that makes it easy to determine how to enter it, and if necessary how to

match it up with the appropriate data recorded by the computer. Figuring this out in advance, and designing your data file and data collection appropriately can save a great deal of time.

2.9 Run analyses with pilot data

We can highly recommend that you run pilot subjects, gather data from them, and analyze the data before launching a large experimental study. The number to run can be found with experience, or by talking with your PI. Analysis of pilot data can provide an approximate baseline of performance, or identify problems with the testing techniques or measures used. Your pilot subjects can be your friends, family, or subjects recruited from your subject pool.

An important aspect of analyzing pilot data is that it provides an opportunity to evaluate your data collection plan. You will learn whether your experimental software is recording the data accurately, or whether pencil-and-paper data are being collected in a way that makes data entry easy. One of us supervised a young researcher who failed to analyze his pilot data, and learned after many hours of data collection that the software he developed was not recording data at all! You will also learn whether the format of your data is appropriate for the analyses you have planned. It is hard to overemphasize the importance of this step in piloting an experiment.

If the results from the pilot data are not what you expected, you can revise the design of the experiment (e.g., change which independent variables are recorded, change the target task, or add another treatments, etc.). If the results from the pilot data match your expectations, you can plan to launch your more formal experiments to gather data to confirm the results. On the other hand, if the pilot results do not match your expectations, they may suggest an interesting new research question.

Keep in mind that with a small number of subjects you might only be able to see large effect sizes. A large effect size means that the difference of your treatment is large with respect to how much people generally vary. For example, freshman will vary in weight, as will seniors, say with a standard deviation of 30 pounds. If the seniors weigh more, like 30 pounds, the effect of going from freshman year to senior year is about the amount the population varies. In this case, the effect size is 30/30 or 1. If, however, these student vary in the number of gray hairs on their heads by 10, and the seniors on average have 1 more gray hair, it will require measuring many more students to show that the number of gray hairs varies than it will take to show that weight varies between these two groups. If you are not finding an effect with a pilot study, you might just need to run more subjects or revise your expected effect size.

2.10 Institutional Review Board (IRB)[9]

Investigators in psychology or human factors in many countries now must obtain approval from the appropriate host institution or organization prior to conducting research. The organization charged with approving research applications in a university setting in the United States is called the *Institutional Review Board* (IRB), which is specific to a university or government lab. The IRB is a committee monitoring, approving, and reviewing biomedical and behavioral research involving humans. The IRB's task is to evaluate the potential risks to subjects (see Chapter 3 for more on potential risks), the compliance of the research with ethical principles and with institutional and government policies, and the suitability of the experimental protocol in protecting subjects and achieving this compliance.

[9] This applies to research in the US. You should enquire locally because some countries do not see risk in routine cognitive experimental projects, or perform reviews in a more local or in a way adjusted more to the type of study.

Before the onset of the experiment, investigators must obtain the informed and voluntary consent of the participants selected for the study. The American Psychological Association's Ethical Principles of Psychologists and Code of Conduct[10] specifies that participants have the right to informed consent—participants have the right to understand what will happen in the study (e.g., any known risks of harm, possible benefits, and other details of the experiment). Only after receiving such a briefing, can a participant agree to take part in the experiment. Thus, the details of the experiment should be written in clear, jargon-free language, and without reference to special technical terms. The participants must be able to easily understand the informed consent form. In addition, the form should enable prospective participants to determine for themselves whether they are willing to participate given his or her situation and personal tolerance for risk. We provide an example of an informed consent form in Appendix 3.

IRB policies are subject to interpretation, so when in doubt contact the IRB representative at your institution. It is useful to think of the IRB staff as coaches, not as police.

In general, IRB reviews fall under two categories, either *expedited* or *full* review. Most behavioral science studies that do not involve the use of experimental drugs, radiation, or medical procedures can be considered for expedited review. Expedited review does not require full IRB approval—that is, the full IRB board does not have to be convened to discuss your study—and an expedited review can usually be accomplished within a few weeks (again this will vary by institution and other factors such as time of year). For all other cases, you will need to go through a full review—these are usually scheduled far in advance at specified dates, and this document does not attempt to cover such studies.

2.11 What needs IRB approval?

Research involving human participants generally requires IRB approval. That sounds simple, but in fact, it is not always easy to decide when you need IRB approval for activities that you consider part of your research. For example, if you or your research assistants participate in your research protocol in the course of developing materials or procedures, IRB approval is not required for your participation for pilot testing; and you cannot publish this data. If, on the other hand, you recruit subjects from a subject pool or the general population for pilot testing or data for publication, you will need IRB approval.

Some other research-like activities that do not require IRB approval include:

- Administrative surveys or interviews for internal use in an organization that will not be published
- Class projects in which only the students in the class provide data and the results will not be published
- Research based on publicly available data

It is easy to confuse this with the "exempt" category established by Federal regulations. This category of research includes research that is truly anonymous (there is no way, even in principle, that participants can be identified) and the research procedures are truly innocuous (cannot cause harm). Examples include the use of standard educational tests in an anonymous fashion, observation of public behavior, or use of publicly-available information.

A complete list of research exempt from IRB review in the US can be found in Title 45, Part 46.101 of the Code of Federal Regulations (http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.101, checked 3 Feb 2012)

[10] http://www.apa.org/ethics/code2002.html

Note that many institutions or funding agencies may require review of research in these categories. For example, Penn State University requires that IRB staff, not the researcher, make the determination that research is exempt.

If you have any questions about whether your project is exempt from IRB approval, you should consult with your IRB, or, if you don't have one, a collaborator or a colleague at a university may be able to provide information. Many IRBs have a special simplified review process to determine whether particular research projects are exempt from review. It is always better to err on the side of caution, and seek IRB approval if in doubt. For example, if you are collecting data in an administrative survey or as part of a class project and *might* want to publish, you should seek IRB approval in advance. The bottom line is that if you are in doubt, you should consult with your local IRB.

Another question arises in research involving collaboration across institutions (e.g., two or more universities, a university and a government agency): Which IRB is responsible for reviewing the research? In general, the answer is that the IRB at the location where data are collected from human participants is responsible. However, this should be established in consultation with your IRB, particularly if you have or are seeking external funding for your research. Some institutions may require that their IRB review the project, regardless of where the data are collected.

If you are working across countries, the U.S. Department of Health and Human Services maintains a compendium of human subjects protections in other countries (www.hhs.gov/ohrp/international/index.html) that may be helpful. Researchers in non-U.S. countries who are unfamiliar with regulations on research with human subjects in their countries may find this a useful starting point.

There are a few other exceptions that are worth noting, where IRB approval is not required. If you are running yourself and only yourself, you do not need IRB approval. If you are running studies only for class work, or for programmatic improvement and not for publication, then IRB is not required. These exceptions are useful when you are piloting studies, or when you are teaching (or learning), or when you are developing software. Of course, you can in most cases still seek IRB approval or advice for situations such as these. The approval process offers you the opportunity for feedback on how to make your study more safe and efficient. Approval also allows later publication if the results are interesting.

IRB approval is required before any aspect of a study intended for publication is performed, including subject recruitment. Without exception, IRB approval cannot be granted once the study has been conducted. Consequently, you should seek IRB approval early in the process and keep your timeline and participant recruitment flexible. You do not need to seek new approval for finishing early or enrolling fewer participants than requested. You will, however, need to seek approval for finishing late or for enrolling a larger number of participants, or otherwise substantially changing the study.

What if you do not have an IRB? For example, you may be conducting behavioral research at a corporation that does not usually do such research. The first question to ask is whether you really do not have an IRB. In the US, if the organization receives Federal government funding, research with human subjects at that organization is subject to Federal regulations. If the organization does not have an "assurance" agreement (an agreement in which the organization offers assurance that they will comply with regulations governing human subjects research) that allows them to operate their own IRB, you should contact the funding agency or the or the Office for Human Research Protections at the U.S. Department of Health and Human Services (http://www.hhs.gov/ohrp/index.html) for guidance on having your research reviewed.

If no Federal funding is involved, as of this writing, there are no legal requirements in the U.S. concerning human subjects research. Of course, you as a researcher still have the same ethical obligations to protect your subjects. The next chapter offers further discussion of potential ethical issues. It is wise in such cases to consult with researchers used to working within IRB guidelines for advice; one of us has sometimes done such consultation with researchers in private industry. And, of course, even if your research is not subject to Federal regulations concerning human subjects, there are still practical reasons for following the guidelines. For example, journals that publish behavioral research generally require that authors certify that their research has been conducted in accord with ethical guidelines. Following accepted practices for the treatment of human subjects may also reduce the risk of legal liability.

2.13 Preparing an IRB submission

New researchers—and experienced ones, too—often find the process of submitting their research to the IRB confusing and frustrating. The details of IRB submission will vary depending on the institution and the nature of the research. We include a sample as an appendix. There are, however, a few things to keep in mind that will help make the process smoother:

- You may first need to get certified yourself. This means, you need to take read some material and pass (typically an online) test showing some basic knowledge about how to run studies and how to treat and protect subjects. Such training is now generally required by IRBs in the United States.
- Many of the questions you will have to answer will seem irrelevant to your research because they are irrelevant—IRB forms must be written to accommodate the wide range of research that must be reviewed, and must include items that allow the IRB members to understand which aspects of the research they must review. For example, this is why you may have to indicate that your study involves no invasive biomedical procedures, when it has nothing to do with any biomedical procedures at all. Also, some of the items may be required by institutional policy or Federal law. Just take the time to understand what is being asked, and patiently answer the questions. Any other approach will just add to your frustration.
- Understand that most of the people involved in reviewing your research will not be experts in your area of research. In fact, by regulation each IRB must contain at least one member of the community who is not associated with the university or organization. This means that it is important to avoid jargon and to explain your procedures in common-sense language. Take the time to write clearly and to proofread—it is to your benefit to make sure that your submission is easy to read. For example, one of us has a colleague who was frustrated that her IRB did not understand that in psychological jargon *affect* means what is commonly called *emotion*—more careful consideration of using common-sense language would have avoided this frustration.
- Get the details right. None of us enjoys filling out forms, but it is especially frustrating to get a form returned to you because some of the details don't match what is expected. One of us had to resubmit an IRB form because of an incorrect email in the personnel list.
- Allow time. Even a perfectly smooth IRB process may take several weeks for an expedited review. A full review may take longer, in part because full reviews are considered at periodic meetings of the full IRB committee. If you must respond to questions asking for clarification, more weeks may be added. If there is disagreement about the acceptability of your protocol, it may take even longer to resolve the situation. Plan accordingly.
- Do what you said you would. While minor changes in your protocol that do not impose greater risks to the subjects generally do not require another IRB review, a modification of

your proposal or any other changes will require another review. For example, if you decide that you want to collect demographic information, add a personality survey, or use a completely different task, consult the IRB staff about how this may affect your approval, and how to get a modification approved.

- Keep good records. IRBs are generally required to conduct occasional laboratory audits on at least a sample of the projects for which they are responsible. If you cannot document your informed consent procedures, show materials consistent with your approved protocol, and so on, the IRB may halt your research while the problems are resolved.
- Ask for help. Find one or more researchers familiar with your local IRB and ask their advice about submitting your study. If possible, find examples of approved protocols to use as models for your own submission. And when in doubt, contact the staff of your IRB with your questions.

2.14 Writing about your experiment before running

It might seem odd to bring up writing in a chapter on preparation for running experiments. On the other hand, writing up your study is the final step, isn't it? That seems obvious to many researchers, and that message is conveyed in many textbooks on research methods. However, it is a good idea to consider writing as part of the preparation process—writing about your hypotheses and their rationales, your methods, even your planned analyses. In some contexts—for example, conducting research for a thesis—researchers are forced to do this. You will never have the details of your thinking about hypotheses, methods, and planned analyses fresher in mind than while you are preparing to run your study. Writing can force you to think through possible gaps in your preparation—for example, if you can't describe how you will manipulate your independent variable, you're probably not ready to actually do it. It may not be useful to spend the time to produce the kind of polished writing you will eventually include in a thesis, a technical report, or a manuscript submitted for publication; but it is useful to think about how you will report to others your research question, your experimental method, and your results.

In particular, writing up your method section before you run your study lets you get feedback on the study before it is run. You can show the method to colleagues and to senior researchers and have them debug the study, pilot it in their minds, before you commit further resources to it. It also means that if you write the method before you run and as you run it will more accurately reflect what you did than if you write it well after the study is completed.

2.15 Preparing to run the low vision HCI study

Studies involving special populations are important but challenging because they by definition involve groups who have different abilities and often need better interfaces and studies using them can be more complex (one example paper starts to take this up Ritter, Kim, Morgan, & Carlson, 2011, but other populations will have other necessary accommodations). Judy's study was no different in this respect. While Judy's study targets a specific special population, blind and partially sighted individuals, we believe outlining the steps and considerations taken in this study will better prepare you for working with other special populations, and, indeed, all study participants.

To conduct her study, Judy and her team had to carefully consider how to best interact with and recruit blind and partially sighted participants; these two considerations are fundamental to studies involving any special populations. The participants in Judy's study differed not only in their visual acuity but also in their opinions regarding blindness and how to best interact with the non-blind world. For instance, experimenters when describing the motivations for the experiment

had to be careful not to assume that the participants viewed blindness as a disadvantage to be overcome; the schism in the deaf community regarding cochlear implants provided a related warning about this effect. Rather, it was more helpful for experimenters to frame in their own minds visual acuity as a characteristic like height that entails a set of attributes and considerations. Further, piloting and preliminary consultations with blind and partially sighted individuals proved crucial for developing a workable experimental plan and procedure.

Visually impaired individuals, like other special populations, are a heterogeneous group. Legal blindness is defined as 20/200 visual acuity or less with glasses or a field of vision less than 20°; however, this general definition masks a whole range of distinctions. Very few visually impaired people have no vision at all or are unable to distinguish light from dark. More generally, partially sighted individuals have blurred vision, restricted vision, or patchy vision. They may have difficulty distinguishing between shapes or colors, or gauging distances. Others may have reduced peripheral vision or conversely good peripheral vision and reduced central vision. Regardless, Judy's experimental plan had to support sighted guiding and room familiarization techniques. In this case, participant recruitment preceded the full development of the experimental plan because achieving a representative sample size depended on the cooperation of outside groups.

Judy's experiment consisted of one *independent variable* (manipulating the navigation bar) and *two treatments* (marking the navigation bar or not marking the navigation bar). The first group (those encountering HTML tags that mark the navigation bar to be skipped unless requested) was the *experimental group*, while the second was the *control group*. The control group used a standard screen-reader that allowed the user to skip to the first non-link line; however, they had to request this action. The experiment's *null hypothesis* was that marking the navigation bar to be skipped unless requested *does not help* blind or partially sighted users. The hypothesis's *dependent variables* were the lag times both within and between viewing the web pages. To effectively establish and test the relationship between these variables, Judy took special care when recruiting participants, preparing the experimenters, and ensuring that the apparatus and test facilities met the participants' needs. We will discuss each of these steps, moving from recruitment to lab setup.

Independently achieving Judy's desired sample size (n=32) outside of a blind institution for 2 sessions was likely to be difficult. Working for a mid-sized company, Judy had to reach out to external groups to find participants. Working with another organization can provide important benefits such as access to experts and assistive technologies; however, such collaborations can also introduce potential logistical, interpersonal, and ethical challenges. We will discuss the potential ethical implications in Chapter 3. For now, we will discuss some of the logistical and interpersonal challenges.

The challenges confronting an experimenter will largely depend on his or her organizational partner. Institutional partners serving students over the age of 18 are likely not only to be able to find participants but also to have resources helpful to the study such as access to facilities, transportation, or orientation and mobility (O&M) specialists. On the other hand, these institutions must protect their students' health and wellbeing, and thus are likely to demand that the study meet the approval of their IRB. Further, you may have to address the concerns of other institutional stakeholders before conducting your study.

If, on the other hand, you turn to an advocacy organization to publicize your study, the degree of institutional support can vary significantly. Support may range from announcements at a local chapter meeting to access to facilities; however, greater support is, again, likely to entail greater institutional oversight, especially if that advocacy organization accepts state or federal funding. When working with an advocacy organization, simply getting the support of its leaders is often

insufficient for achieving a representative sample size. Rather, achieving the level of support necessary to conduct a study frequently requires meeting directly with your potential participants and explaining your study's relevance to them. Also, you may need to provide logistical support in the form of transportation and greeters, as well as compensation. Nevertheless, participants recruited in this fashion are likely to take the study seriously.

Judy partnered with an advocacy organization to find participants. She ran multiple sessions, which made scheduling harder. There are several reasons experimenters use multiple sessions. The most common reason is to study learning or to examine the reliability of the measures. In this case, however, it was because the experiment can't gather enough data in a single session, because it was fatiguing for subjects. In special populations this last reason may be more common.

Because her study did not examine time-sensitive phenomena such as learning or retention, she was able to meet her goals by scheduling two sessions per participant without regard to interval between sessions across two months. If Judy's study had required her to consider the spacing of her sessions, an institutional partner would most likely have been a better match because institutions are better able to provide consistent access to participants. Further, Judy's relatively relaxed time demands enabled her to optimize the study schedule to meet the needs of her participants. Judy did have to work with the advocacy organization to provide clear instructions in multiple formats for reaching her research facility. She also had to ensure that greeters were on hand prior to every session to help conduct participants through the building, and in some cases to meet participants outside of the building.

To effectively greet and work with the study's participants, Judy and her team had to learn both sighted-guiding and room-familiarization techniques. Again, while these techniques are specific to working with partially sighted and blind participants, we include a brief discussion to give some indication of the kind of planning necessary to support studies involving special populations. We do not discuss here related etiquette regarding seeing-eye dogs or participants using mobility tools (e.g., Fishman, 2003; D. S. Kim, Emerson, & Curtis, 2009 for more information). Video tutorials on sighted guiding and room familiarization techniques are available online, including http://www.afb.org/.

Sighted guiding refers to escorting people who are blind or partially sighted through a new or crowded space (Hill & Ponder, 1976). Sighted guiding always begins with the guide asking the person who is blind or partially sighted whether they would like assistance, the participant in this case. Simultaneously, the guide should touch the back of the participant's hand with the back of his or her hand to indicate to the participant his or her relative location (Wisconsin DHS, 2006). The participant and guide should be positioned along the same direction of travel, with the guide half a step in front of the participant. The participant will then grab the participant's elbow before proceeding, with the guide keeping his or her elbow at roughly a right angle. The guide will then describe briefly the room's configuration saying for instance, "We are entering a hallway; or, we are in a large room walking down an aisle with a door ahead of us." The guide will then indicate every time the pair is approaching a door, a curb, a stairway, an obstruction (indicating where the obstruction is in relation to the participant's body), or about to turn. If the participant and guide need to reverse directions, the pair comes to a complete stop with the participant releasing his or her grip. The pair then turns toward each other while executing a 180º turn. The guide then reestablishes contact and positioning before continuing.

Finally, we will discuss setting-up the experimental space in light of the participants' room familiarization needs. Ensuring your experimental space is clean and free of distractions is necessary for preparing any study; however, it takes on special importance in this case. Because participants who are partially sighted or blind will rely on tactile and auditory cues to familiarize

themselves with the experimental space, the lab setup must feature clear walkways (preferably indicated by a different flooring surface or delimited by a boundary), distinct lab spaces with tables pulled slightly away from the walls, and all obstructions (chairs, trashcans, or hanging objects) cleared from not only any walkways but also from the room's perimeter (Marron & Bailey, 1982). Further, the experimenters should clearly explain all auditory cues such as tones indicating the session's start and ending, as well as any other routine but unusual noises that might distract the participant. The lab apparatus should feature clear tactile cues such as buttons that remain depressed while the equipment is in operation; most assistive technologies already include these features but you may find yourself building experimental equipment to investigate your research question.

Although Judy's experimenters helped the participants navigate the lab space, it was important for the participants to be able to navigate the lab without the experimenters' direct assistance, including how to reach the bathrooms. Experimental orientation consisted not only of verbal instructions but also moving around the lab space, starting at the perimeter and then proceeding to the center of the room. During this process, the experimenters indicated important landmarks and answered questions the participants had about the room's layout. Next, the experimenters directed the participants to the experimental workspace and apparatus. All experimental surfaces were kept clear of extraneous materials. As the participant moved from the workspace's perimeter inwards, the experimenter described each piece of apparatus as the participant encountered it, indicating to the participant the key sequences necessary to operate each piece of equipment and answering any question the participant had.

Overall, the preparation steps for this study are the same as other studies. In every case you have to consider how to address your participants, you have to consider how to recruit them and help them arrive safely to your study.

2.16 Preparing to run the HRI study

Human Robotic Interaction[11] (HRI) studies in general require careful preparation because working with robots often requires drawing from multiple skill-sets (e.g., mechanical engineering), and the initial configuration of the study components is not simple or easy to hold consistent and is essential for obtaining meaningful results. To be clearer, robots present researchers with a whole host of issues that can makes small changes in the experimental protocol expensive. So, it is important to try to anticipate problems early, and identify any easy low-cost adjustments if necessary. There are a few examples that can illustrate the issues in this chapter.

Because of additional expenses in time and resources associated with HRI studies, Bob should pilot his study. He may find useful results simply from the piloting work. Taking a risk-driven spiral development approach (Boehm & Hansen, 2001; Pew & Mavor, 2007), he will find many ways to reduce risks to the general success of his company's products, and he will find that even setting-up the study may suggest changes for the robot design related to setting-up the robot repeatedly.

Bob should also prepare his robots and experimental situation carefully. He should try to have the study sessions be otherwise trouble free. Where necessary, there should be backup robots, and enough help or time to reset the task situation back to its initial condition. Just how troublesome the robots are and how long it takes to reset the study situation will become clear from piloting.

When Bob reports his results, he will want his reports to be clear and helpful to the audience for whom he is writing. In some cases, this type of audience has a hard heart, in that they do not

[11] Also sometimes, human-robot interfaces.

want to believe that their products are not usable or user-friendly. If Bob confronts this situation, he should consider not reporting his usability metric at all, but just showing the users' frustration. This is a situation where piloting both study and reporting method may prove essential. Finding the results on paper are not convincing, Bob should consider including a new measure (video tapes of the subjects). Including this measure, however, will require further changes in the protocol, in the apparatus, and in recruitment (notifying subjects that they will be video taped), and in the procedure (getting permissions for the various uses of video tape).

2.17 Conclusion

This is the longest chapter of this book, in part, because most of the work in running an experiment usually goes into preparing to run it. Some of our advice here may seem obsessive, but we have learned from hard experience that cutting corners in preparing to run an experiment usually results not in saving time but in many extra hours repeating experiments or fixing problems. The central point of this chapter is to think carefully about every aspect of the experiment, and to make sure that you have effectively planned as many of the details of the experiment as possible.

2.18 Further readings

We list some reading materials that may help you plan and run experiments, as well as report the results from the experiment.

- Huck, S. W., & Sandler, H. M. (1979). *Rival hypotheses: Alternative interpretations of data based conclusions*. New York, NY: Harper & Row.

Rival hypotheses provides a set of one page mysteries about how data can be interpreted, and what alternative hypotheses might also explain the study's results. Following the mystery is an explanation about what other very plausible rival hypotheses should be considered when interpreting the experiment's results. This book is engaging and teaches critical thinking skills for analyzing experimental data.

- Nielsen, J. (ed.) (1994). Special issue: Usability laboratories. *Behaviour & Information Technology, 13*(1-2).

This is a specially edited article concerning usability laboratories. This special issue provides several representative usability laboratories—mostly computer, telecommunications, and consumer product companies (e.g., IBM, Symantec, SAP, Phillips, or Apple, etc.).

- Ray, W. J., & Slobounov, S. (2006). Fundamentals of EEG methodology in concussion research. In S. M. Slobounov & W. J. Sebastianelli (Eds.), *Foundations of sport-related brain injuries* (pp. 221-240). New York, NY: Springer.

This book chapter provides you with background for using EEG and its processes, including physiological basis and frequency analysis of the EEG. In addition, Ray and Slobounov explain EEG research on motor processes in general and brain trauma specifically.

- Rosson, M. B., & Carroll, J. M. (2002). *Usability engineering: Scenario-based development of human-computer interaction*. San Francisco, CA: Morgan Kaufmann Publishers.

This book provides comprehensive background in the area of human-computer interaction and gathering data about users.

- Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). *Psychophysiological recording* (2nd ed.). New York, NY: Oxford University Press.

Psychophysiological Recording is a very useful book for anyone who conducts experiments with human participants measuring their physiological responses. The book provides not only practical information regarding recording techniques but also the scientific contexts of the techniques.

- Payne, J. W., Braunstein, M. L., & Carroll, J. S. (1978). Exploring predecisional behavior: An alternative approach to decision research. *Organizational Behavior and Human Performance*, *22*, 17-44.

Payne and his colleagues discuss the use of behavioral and verbal protocol approaches to process tracing in the context of decision research. This article illustrates their use of multiple methods to illuminate the decision process.

- Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*, 166-183.

Schooler and his colleagues describe a study in which verbalizing while performing an experimental task changed the nature of the task and interfered with performance. They discuss the circumstances under which such effects are likely to occur. Subsequent research by Schooler and his colleagues is also useful for understanding the potential negative effects of verbalization.

2.19 Questions

Summary questions

1. Answer the following questions.
   (a) What is a "subject pool"?
   (b) What is "verbal protocol analysis"?
   (c) What is required to collect verbal protocol data?
   (d) List several types of dependent measures.
2. What is "error data"? Why is error data expensive to collect?
3. Explain the four types of scales in measurement.

Thought questions

1. Think about a space to run your study. What changes can and should you make to it to improve your study?

2. Search previous studies using the Web of Science or similar bibliographic tool (e.g., Google Scholar, CiteSeer) with a keyword of "speed-accuracy tradeoff". Choose an article that you want and find out what types of independent and dependent measures (e.g., response time, percept correct, etc.) were used in that article.

3 Potential Ethical Problems

Ethical issues arise when the various individuals involved in a situation have different interests and perspectives. Your interests as a researcher may at times be different from the interests of your subjects, colleagues, project sponsors, or the broader scientific community or the general public. People are often surprised when ethical concerns arise in the course of scientific research because they see their own intentions and interests as good. Despite good intentions, however, ethical issues can become ethical problems if they are not considered in the planning and conduct of an experiment. This chapter is concerned with understanding and handling potential ethical concerns.

It is certainly helpful to understand "official" ethical guidelines such as those published by the American Psychological Association (www.apa.org) or those you will encounter in the ethics training required by your university or organization or your professional organization (e.g., The Human Factors Society, the British Psychological Society). The key to making ethical decisions in conducting research, though, is to consider the perspectives of everyone involved in the research process—your subjects, other members of the research team, other members of the scientific or practice community—and to keep in mind the principles of individual choice, honesty, and minimal risk. This is best done before the study—this is not how to fix problems once they occur, but how to avoid problems in the first place.

Ethical concerns can arise at several points in the research process, including recruiting subjects, interacting with subjects during the experiment, handling the data, and reporting the results. We consider each of these stages in turn. All universities and most other organizations have guidelines for research ethics, resources for ethics training, and contacts to discuss ethical issues. And, of course, ethical concerns should be discussed with the lead researcher or principal investigator.

3.1 Preamble: A simple study that hurt somebody

One of us (Ritter) was at another university at a workshop hosted by a non-behavioral science department. In this story the names have been changed to protect all parties. After dinner a good student was brought out to present their undergraduate honors thesis work in computer science as part of the banquet. I think his name was Nidermeyer. He was studying network effects in his cohort. He was the leader of a large group of students, and made them all participate. I looked at my colleague, well trained in human factors, and said, yes? And she said, no, she had no hand in this study.

He then noted that they were given Blackberries to record their movements 24 hours a day. At this institution there are rules about when and where you can be at certain times of day, more so than other institutions. It is highly regimented. I looked at her again, as this movement data could be rather private data, and she just rolled her eyes and said she had nothing to do with this study.

They also gave every participant a survey about their friends in the subgroups, including questions, like, would you have this person date your sister? (It was a nearly but not exclusively a male group). My colleague would no longer look at me or accept questions from me!

Niedermeyer then did the analysis of who was friends with who, creating a social network. In front of the Dean, his thesis supervisor, several teachers in the program (not psychology, thankfully), other students at the college, and the invited guests, Niedermeyer noted that his co-leader in the group, let's call him Bonantz, did not have any friends according to the survey. To understand the results better, he, Niedermeyer, called Bonantz to his room to discuss this result

and to have Bonantz defend why he had no friends. He reported to the room Bonantz' response ("Bonantz did not care").

At this point, that I had seen just about every aspect of good practice violated by this student, and the people nominally in charge of supervising them, including his advisor and the Dean. The student did not take informed consent, he collected and did not protect private data, and he potentially harmed his subject/colleague who was a subject in his study by reporting non-anonymized data.

And as I heard this story, I understood that there was room for experimental ethics education. Maybe I should have stood up and made a piercing comment, but as a visitor, I had little standing, and it would only have hurt Bonantz to emphasize that he was hurt. So, instead, I use this as a teaching story.

In the rest of this chapter, we review the theory in this area, and note some ways to avoid these types of problems.

3.2 The history and role of ethics reviews

We discussed some practical aspects of ethics reviews and working with your Institutional Review Board (IRB) in the previous chapter. Before discussing the ethical concerns you may have to consider in developing your research, it is useful to briefly consider the history of ethics reviews of research with human subjects. Much of currently accepted practice with respect to the ethics of behavioral research results from concerns with medical and behavioral research in the past century. Many of the regulations governing research with human subjects in the United States grew out of controversial medical research (as can be contrasted with behavioral research). An overview of the history is available from the US Department of Health and Human Services (www.hhs.gov/ohrp/archive/irb/irb_introduction.htm). Beginning in 1966, the National Institutes of Health issued guidelines that established Institutional Review Boards (IRBs) as a mechanism for reviewing research with human subjects.

The most direct influence on current regulations was the reaction to the Tuskegee Syphilis Study, in which the U.S. Public Health Service monitored the progression of syphilis in hundreds of African-American men, while failing to offer treatment even after a known cure (penicillin) became available. When this study was revealed in the early 1970s, the United States Congress passed legislation creating the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. This legislation, the National Research Act, began the process by which IRB review became mandatory for behavioral research with human subjects.

The Commission issued a number of reports, the last of which is known as the Belmont Report (http://ohsr.od.nih.gov/guidelines/belmont.html). This report provided guidelines for research with human subjects, based on the principles of respect for persons, beneficence, and justice. It lays out the basic guidelines for informed consent and assessment of the risks and benefits of participating in research. The report is quite brief, and is well worth reading to understand the background of the review process conducted by your IRB.

Currently (as of 2011), oversight of human subjects research is the responsibility of the Office of Human Subjects Research (http://ohsr.od.nih.gov/), which is part of the National Institutes of Health. This office oversees the operation of IRBs operating at universities and colleges.

3.3 Recruiting subjects

Ethical concerns with respect to your subjects begin with the recruiting process. Obviously, you should be honest in describing your study in your recruiting materials, including accurate

statements of the time and activity required and the compensation provided. Perhaps less obvious, it is important to think about fairness with regards to the opportunity to participate. For example, if you are using a university subject pool, you will have to justify scientifically any criteria that might exclude some students.

Usually, we would like to generalize the results that we find to a wide population, indeed, the whole population. It is useful to recruit a representative population of subjects to accomplish this. It has been noted by some observers that experimenters do not always recruit from the whole population. In some studies, this is a justifiable approach to ensure reliability (for example, using a single sex in a hormonal study) or to protect subjects who are at greater risk because of the study (for example, non-caffeine users in a caffeine study).

Where there are no threats to validity, however, experimenters should take some care to include a representative population. This may mean putting up posters outside of your department, and it may include paying attention to the study's sex balance or even age balance, correcting imbalances where necessary by recruiting more subjects with these features.

As the research assistant, you can be the first to notice this, bring it to the attention of the investigator, and thus help to address the issue.

3.4 Coercion of participants

You should not include any procedures in a study that restrict participants' freedom of consent regarding their participation in a study. Some participants, including minors, patients, prisoners, and individuals who are cognitively impaired are more vulnerable to coercion. For example, enticed by the possibility of payments, minors might ask to participate in a study. If, however, they do so without parental consent, this is unethical because they are not old enough to give their consent—agreements by a minor are not legally binding.

Students are also vulnerable to exploitation. The grade economy presents difficulties, particularly for classes where a lab component is integrated into the curriculum. In these cases, professors must not only offer an experiment relevant to the students' coursework but also offer alternatives to participating in the experiment.

To address these problems, it is necessary to identify potential conditions that would compromise the participants' freedom of choice. For instance, in the example class with a lab component, recall that it was necessary for the professor to provide an alternative way to obtain credit. In addition, this means ensuring that no other form of social coercion has influenced the participants' choice to engage in the study. Teasing, taunts, jokes, inappropriate comments, or implicit quid pro quo arrangements (for example, a teacher implies that participating in their study pool study will help students in a class) are all inappropriate. These interactions can lead to hard feelings (that's why they are ethical problems!), and loss of good will towards experiments in general and you and your lab in particular.

3.5 Risks, costs, and benefits of participation

Most research participation poses little risk to subjects—a common phrase is "no risks beyond those encountered in everyday life." However, participating in research does carry a cost for the subject; he or she devotes her time and effort to getting to the experimental location and performing the task. Of course, subjects benefit when they are compensated by money or course credit, or even by the knowledge that they have contributed to an important piece of research. Nevertheless, ethical guidelines for human subjects require that the researcher weigh the benefits of the research—the value of the data collected, the compensation the subject receives—against whatever costs and risks the subject may encounter. It is common for university subject pools to

require that subjects benefit not just by receiving course credit for participating but also by learning something about the research topic and process, usually through a debriefing at the end of the experiment.

Sometimes, there are physical or psychological risks beyond those encountered in everyday life. Even very simple procedures such as attaching electrodes for electrophysiological recording have some risk, as do experimental manipulations such as asking subjects to consume caffeine or sweetened drinks (some people are sensitive to caffeine, some are diabetic). It is important to consider these risks.

More common than physical risks are psychological risks. The collection of sensitive data, which we discuss, next carries risks, as do experiments featuring deception or procedures such as mood induction. When considering procedures that involve psychological risks, it is important to ask whether the procedures are essential for scientific reasons—deception, for example, often is not—and whether the benefits outweigh those risks. Often, it is important to withhold such information as the nature of your research hypotheses because you want to study your subjects' natural behavior in the experimental task, not their effort to comply (or not comply) with your expectations. This is not deception because you can withhold this information while truthfully informing subjects about what they will experience in your experiment.

Another example of psychological risk is stress. Stress can result from experimental tasks that place high cognitive demands on subjects, from the conditions in the laboratory (e.g., heat, noise), or social pressure on the subjects. Sometimes, stress may be manipulated as an independent variable, as in studies of the effect of time pressure or social threat (what someone might think of a subject's performance) on mental processes. It is important to minimize sources of stress that are not relevant to the study, and to monitor subjects' reactions to stressors that must be included. In some cases, it may be necessary to halt an experimental session if a subject is becoming too stressed. If stress is included as part of the experimental design, your debriefing should address the need to include it and allow you to address the subjects' reactions.

While it is common to think about risks only in terms of the research procedures, there is another category of risks that should be considered. For example, if you conduct research on a university campus, especially in an urban setting, the time of day at which you hold experimental sessions may pose risks to participants. Students leaving your lab after dark may be at risk simply by walking unaccompanied in the area. This may seem like an everyday risk that has nothing to do with your experiment, but it is something to consider—students have been accosted leaving labs, and it is useful to think about this possibility in planning your study.

3.6 Sensitive data

When preparing to run a study, you should consider how you will handle sensitive data. Sensitive data include information that could violate a subject's privacy, cause embarrassment to a subject, put the subject at risk of legal action, or reveal a physical or psychological risk previously unknown to the subject. Your research question may require that you collect data that you anticipate will be sensitive, or you may collect data that is unexpectedly sensitive. While all data collected from human subjects is generally considered confidential—that is, not to be shared—sensitive data requires additional precautions.

The most common kind of sensitive data is personal information. Such information includes an individual's race, creed, gender, gender preference, religion, friendships, income, and so on. Such data may be important to your research question; for example, you may be interested in whether the effects of your independent variable depend on membership in certain demographic categories. Similarly, you may want to collect personal information using questionnaires

designed to assess personality, intelligence, or other psychological characteristics. However, when such data can be associated with individuals' names or other identifying information, a risk of violating privacy is created. These data should not be shared with people not working on the project, either formally if you have an IRB that requires notice, or informally, if your IRB does not have this provision (this may occur more often outside of the US). You should seek advice from your colleagues about what practices are appropriate in your specific context.

A second type of sensitive data involves subjects' responses have implications outside of the scope of the study. For example, some research questions require data about topics such as the use of recreational drugs, tobacco or alcohol use, or medical conditions. For example, if you are administering caffeine, and you ask the subject what drugs they take (to avoid known caffeine agonists or antagonists), you may find information about illegal drug use. Such data can pose a variety of risks: legal risks if subjects reveal illegal activity, risks to employment status or insurance eligibility, and so on. Data received from other sources may also contain sensitive data. In one recent case, a research sponsor provided database records concerning subjects in a study, and a researcher posted parts of the data on the Internet without realizing that the records included social security numbers! Obviously, greater care would have avoided risks to these subjects.

These kinds of sensitive data can be anticipated, and precautions beyond the routine protections of confidentiality can be planned. For example, you may collect and store the data in such a way that it cannot be associated with a subject's identity, instead using identifying codes (subject IDs) to link the various components of a subject's data. Removing identifying information from a data set is sometimes referred to as *anonymizing* the data. However, it is important to be aware that removing identity information may not be sufficient to successfully anonymize data, if you have collected demographic information. For example, one study showed that knowing the 5-digit zip code, gender, and date of birth is sufficient to identify 87% of Americans (Sweeney, 2000). In smaller samples, known basketball players or certified soccer refs in a lab will uniquely distinguish many people. The measures you need to take will depend on the nature of the data you collect, and how they will be stored and shared. Your local IRB or ethics review board, as well as experienced researchers, can provide guidance on standard practices. Under some circumstances, researchers in the US can protect sensitive data by requesting Certificates of Confidentiality from the National Institutes of Health (NIH), which "allow the investigator and others who have access to research records to refuse to disclose identifying information in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level" (http://grants.nih.gov/grants/policy/coc/background.htm). The web site provides more detail on the circumstances and limits on these Certificates.

IRBs routinely require that sensitive data—and in some cases any data—be stored on secure, password-protected computer systems or in locked cabinets. Extra precautions may include storing the data only on computer systems in the research facility or using the data only in the laboratory, rather than carrying it on a laptop computer or portable storage device. When sensitive data must be shared among members of a research team, perhaps at multiple locations, it is important to arrange secure transport of the data. For example, sensitive data generally should not be transmitted by email attachments.

Data are usually reported in the aggregate, but sometimes you may want to discuss subject-specific responses in writing about your data. For example, in studies using verbal protocols, it is common to quote parts of specific protocols. Skill acquisition studies sometimes display learning curves for individual subjects. Such references to individual subjects should be made anonymous by using codes such as subject numbers rather than potentially identifying information such as first names or initials.

Sometimes, conducting research can lead to unexpectedly finding sensitive data. For example, commonly-used questionnaires for assessing personality or mood may reveal that a subject is suicidal. Or taking a subject's heart rate or blood pressure measurements may uncover symptoms of underlying disease. In such cases, a researcher is ethically obligated to take action, such as referring a subject to appropriate resources. Guidance on these actions is often available from your IRB or ethics panel. The central point, though, is to be prepared for sensitive data and understand how to address such situations.

3.7 Plagiarism

Plagiarism refers to taking other's work or ideas and using them as one's own, that is, without attribution. Particularly in academia, this problem is taken seriously.

An individual might be tempted to steal others' ideas, research methods, or results from unpublished or published works. Nowadays, manuscripts that are about to be submitted or already submitted for review, can be available online.

Why are people tempted to plagiarize others' work? Generally, pressure to meet or surpass institutional standards causes people to plagiarize. To pass a programming class, students might copy another student's code. A faculty member, facing review for tenure and stressed by the number of his or her refereed publications, or an RA trying to fill in a methods section all might be tempted to steal the work of others. Sometimes, the pressure to publish is enough to tempt an academic to plagiarize other's ideas and fabricate their data.

The integrity and development of scientific knowledge is rooted in the proper attribution of credit. In the APA's publication manual (p. 349), you can find the APA's guidelines for giving credit. Direct quotes require quotation marks and citations while paraphrasing or in anyway borrowing from the work of others requires a citation. You may also need to acknowledge people who give you unpublished ideas for your research designs. In particular, you may have personal communications (e.g., email, messages from discussion groups on the net, letters, memos, etc.) that require acknowledgement. In this case, you will need to remember who gave you the idea (an email thanking them can be a good way to document this), and then cite them in the text with a date.

3.8 Fraud

We, sometimes, are shocked by news about research fraud. For example, if a researcher fabricates data and publishes a paper with the data, this is fraud. Other scientists trying to replicate the results are often the ones who find and reveal the initial findings to be fraudulent. While research fraud is unusual, we, nevertheless, must be aware that fraud can cause significant adverse effects not only for the perpetrator of the fraud but also often second or third parties such as his or her academic colleagues, institution, funding agency, or corresponding journal editor. Fraud can also affect more distant people who base key choices on the work in question (e.g., an educational system that prioritizes curriculum strategies based on fraudulent learning data).

If data is lost, it is lost; do not replace it. If you accidentally delete data, do not replace it. If you did not run a subject, do not run yourself. All of these practices undermine your study's validity and are extremely egregious ethical violations. It is sad when you read in an article that "data from 3 subjects were lost", but it is far better to write this phrase than to commit fraud.

3.9 Conflicts of interest

Conflicts of interest arise when non-scientific interests are at odds with the goal of doing good, objective research. These non-scientific interests are often financial—a researcher may be aware of what conclusions a research sponsor would like to see from the research. Conflicts of interest can also be local. For example, a research assistant may know his or her supervisor's favorite hypothesis, and which data would support that hypothesis. Conflicts of interest can, of course, lead to outright fraud, as when a researcher fabricates results to please a sponsor. More commonly, conflicts of interest can influence—even unwittingly—the many scientific judgment calls involved in conducting research. For example, deciding that a subject did not follow instructions and thus should be excluded from the data analysis is an important decision. It is imperative that such decisions do not occur simply because the subject in question did not provide data that fits a favorite hypothesis.

Long term, quality people at quality institutions working with quality theories grapple with these conflicts in a civil, productive, and routine way. This is how science moves forward. Sometimes the unexpected data leads to drastically new and useful theories, sometimes surprising data leads to questions about how the data was gathered and how well the apparatus was working that day. These discussions of interpretation and measurement are normal and you should participate in them appropriately and mindful of the issues.

3.10 Authorship and data ownership

Most behavioral research involves extensive work by multiple individuals, and these individuals should receive appropriate credit. "Appropriate credit" often means acknowledgement in published reports of the research. Such acknowledgement may take the form of authorship, or thanks expressed in a footnote. Assigning credit can raise ethical issues because the individuals involved may disagree about how much credit each member of the team should receive and how that should be acknowledged. According to the American Psychological Associations code of ethics (http://www.apa.org/ethics/code/index.aspx)

> Principal authorship and other publication credits accurately reflect the relative scientific or professional contributions of the individuals involved, regardless of their relative status. Mere possession of an institutional position, such as department chair, does not justify authorship credit. Minor contributions to the research or to the writing for publications are acknowledged appropriately, such as in footnotes or in an introductory statement.
> (APA Ethical Principles, 8.12 (b))

While this sounds straightforward, it leaves room for disagreement, partly because each individual is most aware of their own contributions. The most useful way to address this is to talk about it, preferably early in the research process. Note that simply performing work in the lab under the direction of someone else does not necessarily constitute a scientific contribution. Useful discussions but not complete answers are available (e.g., Darley, Zanna, & Roediger, 2003, p. 122-124; Digiusto, 1994).

A related issue that sometimes arises is about data ownership. That is, who has the right to decide to share the data, or to use it for other purposes? Generally speaking, it is the principal investigator who owns the data, but many other considerations can come into play. In some cases, the data may be proprietary due to the sponsorship arrangement.

It is easy for disagreements about data ownership to arise. For example, does a collaborator who disagrees with the principal investigator about the conclusions to be drawn from the data have a right to separately publish his or her analyses and conclusions? Does a student working as part of

research team have the right to use the data collected by that team for other purposes, such as additional analyses? May a student leaving a lab (such as a graduate student leaving for an academic job) take copies of data collected in the lab? May a student send the data from a conference presentation to someone with whom he or she discussed the project at a conference? How do the answers to these questions depend on the scientific contribution of the student? Note that data ownership has implications for who may benefit from access to the data (e.g., by publishing the results), and for who has responsibility for archiving the data, protecting its confidentiality, and so on. Again, the most useful way to address this issue is to discuss it openly and clearly. Simply relying on future collegiality or unspoken assumptions is likely to and has routinely resulted in problems.

3.11 Potential ethical problems in the low vision HCI study

Studies involving special populations are important because their findings, whether contributing to new technologies or informing public policy, can and have had a lasting effect on the lives of those of individuals. These effects include the development of assistive technologies, as well as results to support the adoption of beneficial legislation such as the American with Disabilities Act (ADA). These effects, however, also include products and policies that have led to the profound mistreatment of numerous vulnerable groups. In light of this history, it is essential that experimenters not only enforce informed consent procedures but also ensure that participants feel that the experiment is in no way a personal assessment. During Judy's pilot experiments, she found that it was important for the experimenters to explain to the participants that they were not necessarily expected to perform the experimental task to some standard or even to complete it; this diffused tension and encouraged the participants who had difficulty completing the task to explain why, which was a very useful result.

As noted in Chapter 2, collaborating with outside groups can introduce both logistical and ethical problems. In Judy's case, her organizational partner was an advocacy organization made up of self-sufficient adult volunteers, some who were blind and other who were not. There were instances where Judy and her organizational partner did assist with transportation; however, the participants in all these cases could decline to participate at any time. Participants participating in the study could choose to either go directly to Judy's office or to meet volunteers at the organization's center, where they would receive a ride to Judy's office. Participants who were unable for any reason to make it to either Judy's office or the center were not contacted again regarding their involvement in the study. In the event that a participant did contact either Judy or the volunteers, a new appointment was scheduled without recrimination. In most instances, we would advice experimenters to avoid re-scheduling trials and instead find more participants. In this case, however, Judy's pool of participants was relatively small and there were few instances where this was an issue.

In the one case where a participant was unable to make it to a trial twice in a row, when they contact Judy again, she discussed with them about rescheduling, but letting them know that they are not obligated to do so. This is a delicate situation, balanced between being sympathetic to difficulties in getting to the study and health problems, supporting subjects who have taken on participation in the study as an obligation they would like to fulfill, to releasing subjects who cannot make it or whose ability to participate has declined.

Compensation is another source of potential ethical challenges, particularly for special populations. Compensation can be a form of coercion if it masks an essentially mandatory act. With regards to special populations, this most often occurs when the participants' freedom of choice is effectively compromised by their status as members of that group. When working outside of an institution that would provide monitoring, like a school for the blind, experimenters are most likely to confront this situation when a participant's circumstances force him or her to

view participating in a study as more than a meaningful use of time. With unemployment rates for persons who are blind estimated to be between sixty and seventy percent, this was a real concern for Judy (AFB, 2012). This statistic is not a direct indicator of quality-of-life however; family support, savings, pensions, and social services are important factors. Consequently, when weighing the potential risk of unduly influencing a participant to enroll in your study, a holistic assessment is necessary. Also, it is important to set a fair rate of compensation, generally ten to fifteen percent higher than the population of interest's median income. This heuristic is only a rule-of-thumb, but it does generally provide your participants an attractive but not overwhelming option.

Finally, while Judy did not work with an institutional partner, we should briefly discuss recruiting participants from such organizations. As noted earlier, some research questions require routine access to larger subject pools to investigate; institutions are often the de facto choice to find pool of subject. When looking for an institutional partner that will provide access to participants (as opposed to expert advice or technical assistance), we advise choosing institutions who have an established IRB. While these organizations may insist on working through their review process, the risk of recruiting participants who are coerced into participating is less. Also, these institutions are already familiar with the experimental process, and thus are more likely to be better equipped to support a behavioral study.

3.12 Potential ethical problems in the multilingual fonts study

Ethics is right conduct. Right conduct interacts with and informs many of the practicalities of running a study, protecting the validity of a study, and also protecting the rights and comfort of subjects. This example examines some of the places where these topics interact.

Research in many academic settings depends on the routine collaboration of more and less experienced researchers. These collaborations, while often fruitful and potentially life changing, can present interpersonal and sometimes ethical challenges to both parties. While these challenges can arise from cultural differences, they are often the byproduct of a disparity between the collaborators' research skills and their managerial skills. Cultural differences can intensify the effects of this disparity, but they can also be conflated with them.

While gaining greater technical skills over time frequently entails acquiring more managerial skills, these are two distinct skill sets that require both thought and practice. Consequently, we believe mitigating interpersonal conflicts associated with routine student collaborations requires PIs, lab managers, and collaborating students to address not only cultural sensitivity issues but also basic communication and project management skills. We will try to develop this guidance a bit further by discussing Edward's and Ying's experiment.

While Ying had felt that the first meetings with Edward went well, over time she became frustrated with Edward's performance. Specifically, Edward had arrived late to run a pilot subject, had sporadically forgotten to either post or take down the "Running Subjects" sign, and had on two occasions failed to backup data on the lab's external hard drive. With these basic lapses, Ying became increasing concerned that Edward was making other mistakes. When she had brought these issues to his attention, he, at first, seemed to earnestly try to correct his mistakes. Later, however, he just appeared frustrated, saying that he had it under control. Well, Ying wasn't so sure that Edward did have it under control, leading her to speak with the PI regarding Edward. While she was busy analyzing the pilot data, Ying knew she had a problem, but felt time pressured herself and thus becomes increasingly frustrated and angry with Edward.

Edward believed Ying was basically nice, but also busy, difficult to understand, and somewhat aloof. Feeling bad about being late and forgetting to backup experimental data, Edward felt that a lot of his issues with Ying were the result of poor communication. Nevertheless, he felt awkward

asking questions because he did want to appear to be emphasizing Ying's difficulty with certain English words. Also, Ying's reactions to some of his questions had made him feel stupid on occasion, as if everyone but him was born knowing how to run an experiment. For instance, it wasn't until his weekly meeting with the PI that he really understood why leaving the "Running Subjects" sign up is a big deal, or where to check for experiment times. Ying briefed him on the experimental protocol in detail, but never mentioned where the weekly schedule was located. In fact, Ying would always tell him the day before the study. He was late to a pilot session only after missing a day in the lab due to an illness. Edward thought Ying would call him and let him know if anything important was coming up. Edward only found out about the pilot session because a friend in the lab had called him. As for backing-up the data, Edward often found himself rushing at the end of the day because the last bus to his apartment complex left shortly after the last experimental session. He hadn't told Ying because he felt it shouldn't be her problem. So, in his rush to catch the bus, he had twice forgotten to back-up the data. To correct these oversights, Edward wrote, "backing up data", as an additional step on the experimental protocol that Ying gave him to help him remember. After writing this last step down, Edward has not failed to backup the data. Nevertheless, Ying was still clearly concerned about his performance, but hesitant to directly address the issue. Instead, she expressed her concern through hyper vigilance. All Ying's double-checking made Edward resentful which in turn made him less focused at work.

As in Edward's and Ying's case, most interpersonal conflicts between collaborating students do not arise from an initial lack of good will but rather an incomplete theory of mind, a psychological term that refers to our assumptions about what those around us believe, know, and intend. Until definitively proven otherwise, you should assume that everyone involved in the collaboration really does want to make it work. When we become frustrated or stressed, we can fall back on generalizations to explain why there appears to be gaps in our understanding and that of our colleagues. In some instances, these generalizations can have an element of truth. These generalizations, however, rarely lead to improved understanding or a stronger collaboration. Rather, stronger collaborations arise from a concerted organized effort to establish a common theory of mind. In smaller labs, the coordination necessary between researchers is smaller and easier to maintain, and the researchers know each other relatively well. These informal interactions can lead experimenters to believe that a common theory of mind just emerges because many of the practices that support it are performed by the PI, and thus in some senses are invisible to the rest of the lab. The quality of these processes is often what determines the lab's ability to instill the problem-solving skills necessary for clear experimental insights; alienated students are generally not critical thinkers or observers.

As experimenters begin to design and run their own experiments, they begin to inherit these coordinating duties. Often, lab protocols and procedures make this process easier. On the other hand, protocols and procedures cannot envision every possibility and generally do not operate at such a fine level of granularity as to encompass many of these basic coordinating functions. At the onset of an experiment, a Ph.D. candidate should ask for the basic contact information of any junior RAs assigned to assist him or her, review expectations, and discuss any limitations. Think broadly, limitations could encompass unknown disabilities or health issues, scheduling problems, or other situational factors. Simultaneously, be courteous when asking situational questions and limit your questions to what you need know to help your junior colleague succeed in the lab.

To better envision the needs of your junior colleagues, ask yourself, "What do I think this student needs to know, and what do I think this student already knows?" Crucially, we need to test our hypotheses by asking junior RAs pertinent questions in a comfortable setting at the beginning of the experimental process, as well as checking with your advisor to verify that you have fully accounted for the information the RAs will need to know to complete their specific experimental

tasks. Try to meet with each RA one-on-one if possible. In these meetings, record the RAs' responses. Next, assess each RA's strengths and weaknesses, see if there are any trends, and devise reinforcement strategies that meet the needs of each RA supporting your experiment. These strategies may be collective strategies such as checklists or specific strategies such as assigning personalized readings or conducting individualized training. Group rehearsals are another important way to anticipate the needs of your junior colleagues. Walk through the entire experimental process from setup to tear-down with your RAs, and amend written protocols in response to questions or missteps that occur during these rehearsals.

For Ph.D. candidates (and PIs), we also suggest that you check-in with the students assisting you in your experiments. You can do this in several ways. First, greet your team members and lab mates, not every interaction with a junior colleague should begin with a problem. Taking a moment or two to greet, thank, or encourage your colleagues can go a long way towards lab relations. Second, have set times to meet with junior colleagues to discuss the experiment and address problems. If a problem has been identified and it does not pose a direct risk to anyone involved in the experiment or its success, ask the RA how he or she might remedy it before interjecting your own answer. The student may have an innovative idea, and in either case, you are allowing the RA to take ownership of his or her work process.

If you are working with multiple RAs, make sure you address any problems in performance privately—the point is not to shame the RA into compliance. Also, try to resolve problems at the lowest level possible; this not only encourages a sense of trust but also makes instances where you do need to call in the PI symbolically more significant. In other words, your colleagues will immediately understand that any instance where the PI is asked to intervene is significant, and thus is to be taken seriously. Finally, distinguish between checking-in and micro-managing. Touching base with your team at the beginning of the day, during specific delicate steps, and at the end of the day will allow you to maintain not only situation awareness but also the perception that you value your colleagues' input and time. Otherwise, unless they are making obvious or potentially dangerous mistakes, let your team members do their jobs.

While Ph.D. candidates have obligations to their junior colleagues, incoming RAs, whether graduate or undergraduate students, also have an obligation to communicate and engage. A successful internship is not a passive affair. Like you, your senior colleagues are mortals who must operate under imperfect conditions based on incomplete information. From the beginning of your internship, try to envision what you, your colleagues, and your supervisors need to know to succeed. Throughout your time in the lab, we suggest periodically returning to this question. Identifying scheduling conflicts or other basic conditions for success does not necessarily entail technical expertise. On the other hand, as you gain technical expertise, return to this question. Better anticipating the needs of your colleagues and supervisors, whether in a lab setting or elsewhere, is strongly correlated with success. Also, asking for clarification or reasons for a particular step is important. If framed within the scope of the project, these questions are unlikely to cause offense. Further, communicating what steps you have taken to resolve a problem, even if imperfect, builds good will and indicates a reassuring seriousness of purpose. In the case of Edward and Ying, there collaborative challenges were the result of both parties failing to interrogate their assumptions about the needs and knowledge of the other.

Returning to Edward and Ying, they were able to resolve their difficulties. The PI, observing in her weekly meetings a breakdown in communication, pulled Edward and Ying into her office and worked through the initial issues that led to the problems in communication and performance. Sometimes, an arbiter is necessary. This session alone; however, was not sufficient to build a strong collaboration. Edward and Ying through several weeks and many clarifications were able to build a routine that worked. This would not have been possible if each had not trusted that at a very basic level the other wanted the team and each other to succeed. As communication

improved, performance improved (to a great extent because Edward better understood what Ying was asking him to look for, anticipate, and do), and gradually the team gained more confidence and momentum.

3.13 Conclusion

As this chapter shows, there can be a large number of ethical considerations involved in running an experiment. Depending on your role in the research, some of them—for example, authorship and data ownership—may be someone else's responsibility. Nevertheless, everyone who participates in the development and running of an experiment must be aware of the possible ethical problems, knowledgeable about the relevant principles and policies, and sensitive to how subjects are treated both while they are in the lab and while the data they provide is being stored, analyzed, and reported.

This chapter notes a few of the most important ethical problems you might face. You may encounter others. If you have questions, you should contact the lead investigator or other senior personnel. In some cases, as in many ethical situations, there may not be a right answer—there may be several right answers. Often, however, there are better answers and good accepted practices.

3.14 Further readings

Here is a list of further readings for you concerning this chapter.

- The APA's webpage, *Ethical Principles of Psychologists and Code of Conduct*. This was published first in 1992, but has been superseded by newer releases (http://www.apa.org/ethics/code/index.aspx).
- American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

  The APA publication manual provides useful guidance for reporting your experimental findings in a written paper.
- Singer, J. A., & Vinson, N. G. (2002). Ethical issues in empirical studies of software engineering. *IEEE Transactions On Software Engineering 28*, 1171-1180.

  This article provides practical advice about ethical issues in running studies in software engineering. In doing so, it provides a resource that would be useful in many similar studies, e.g., HCI, systems engineering, and other situations studying work in companies.

3.15 Questions

Summary questions

1. Answer the following questions.

   (a) What are sensitive data? Give several examples that you will not typically see, and give several examples that you might see in your area.

   (b) What is "plagiarism"?

   (c) What is "counter balancing"?

Thought questions

1. Discuss the way that you can anonymize the sensitive data in your experiment.

2. Recall the question number 2 in Thought Questions, Chapter 1 (the operational definitions of the research variables). Suppose that you will conduct a research study with these variables. Discuss how to plan "recruiting subjects" with consideration of ethical concerns (i.e., how to explain your study to subjects, what is the inclusion and exclusion criteria of the subject, how to use a subject pool, how to protect subjects if there is any known risks, etc.)

3. For each of the major concerns in this chapter (as noted by section headings), note a potential concern for each of the running examples (X study, Y study, HRI study).

4 Risks to Validity to Avoid While Running an Experiment

Understanding how subjects will complete the task, and working towards uniformity across all iterations of the procedure for each subject are important. The repeatability of the experiment is a necessary condition for scientific validity. There are, however, several well-known effects that can affect the experimental process. Chief among these are experimenter's effects, or the influence of the experimenter's presence on the participants and how this effect can vary across experimenters. Depending upon the experimental context and the experimenter, the experimenter effects can lead to either better or decreased performance or a greater or lesser effect of the IVs on the DVs. The magnitude and type of effect that can be attributed to the influence of the experimenter generally depends upon the type and extent of personal interaction between the participant and experimenter. Thus, you should strive to provide each participant the same comfortable but neutral testing experience.

Besides experimenter effects, there are other risks to the experimental process. We highlight some here and illustrate how to avoid them, either directly or through proper randomization. Randomization is particularly important because you will most likely be responsible for implementing treatments. Understanding other risks to validity, however, will also help you take steps to minimize biases in your data. Finally, there are other experimental effects that are outside of your control—we do not cover all of these here (for example, the effect of historical events on your study). Even though you cannot eliminate all contingent events, you can note idiosyncrasies, and with the principle investigator either correct them or report them as a potential problem.

Another common source of variation across trials is the effect of the experimental equipment. For instance, if you are having subjects interact with a computer or other fixed display, you should take modest steps to make sure that the participant's distance to the display is the same for each subject—this does not mean, necessarily, putting up a tape measure, but, in some cases, it does. It is necessary to be aware that the viewing distance can influence performance and in extreme cases can affect vision, irritate eyes, cause headaches, and change the movement of the torso and head (e.g., Rempel, Willms, Anshel, Jaschinski, & Sheedy, 2007). Because viewing distance influences behavior, this factor can be a risk to validity. Furthermore, if subjects are picking up blocks or cards or other objects, the objects should be either always in the same positions, or they should be always randomly placed because some puzzle layouts can make the puzzles much easier to solve (e.g., Jones, Ritter, & Wood, 2000). The experimental set up should not be sometimes one configuration and at other times another.

There will be other effects where variation in the apparatus can lead to unintended differences, and you should take advice locally to learn how to reduce them.

4.1 Validity defined: Surface, internal, and external

We refer to validity as the degree to which an experiment leads to an intended conclusion from the data. In general, two types of validity, internal validity and external validity, are of interest. Internal validity refers to how well experimental treatments explain the outcomes from the experiment. The experimental treatments indicate independent variables that you design. External validity, in contrast, refers to how well the outcomes from the experiment explain the phenomena outside the designed experiment. This is known as "generalizability".

Campbell and Stanley (1963) discusses 12 factors that endanger the internal and external validity of experiments. We need to consider how to reduce or eliminate the effects associated with these factors to guarantee valid results.

When you run studies you may notice factors that can influence the ability of the study results to be explained (this is referred to as "internal validity"). Because you are running the subjects, you have a particular and in many ways not repeatable chance to see these factors in action. Good principle investigators will appreciate you bringing these problems to their attention. You should not panic—some of these are inevitable in some study formats; but if they are unanticipated or large, then they may be interesting or the study may need to be modified to avoid them.

History: Besides the experimental variable, a specific event could occur between the first and second measurements. Typically, this is some news item such as a space launch or a disaster that influences subjects in a global way leading to better or worse results than would occur at other times. But local events like a big football game weekend can also cause such changes.

Maturation: Participants can grow older, become more knowledgeable, or become more tired with the passage of the time. Thus, if you measure students at the beginning of the school year and then months later, they may get better scores based on their having taken classes.

Testing: The effects of taking a test on the scores of a second test. For instance, if you take an IQ test or a working memory test and then take the same test a second time, you are likely to score better, particularly if you got feedback from the first taking.

Instrumentation: It is required to calibrate a measuring instrument regularly. Some instruments need to be recalibrated with changes in humidity. Failure to recalibrate can affect an experiment's results.

Statistical regression: We need to avoid selecting groups on the basis of their extreme scores. If you select subjects based on a high score, some of those high scores will most likely not reflect the participant's normal performance, but a relatively high score. On retests, their performance will decrease not because of the manipulation but because the 2^{nd} measure is less likely to be extreme again.

Selection Biases: Differential selection of participants for the comparison groups should be avoided. Subjects that come early in the semester to get paid or get course credit are different from the subjects who put it off until the last week of the semester.

Experimental mortality: There could be a differential loss of participants from the comparison groups in a multi-session study. Some conditions could be harder on the subjects, and thus lead them to come back less in a multi-session study.

As you run subjects, you may also see factors that influence the ability to generalize the results of the study to other situations. The ability of results to generalize to other situations is referred to as external validity.

The reactive or interaction effect of testing: A pretest could affect (increase or decrease) the participants' sensitivity or responsiveness to the experimental variable. Some pre-tests disclose what the study is designed to study. If the pre-test asks about time spent studying math and playing math games, you can bet that mathematical reasoning is being studied in the experiment.

The interaction effects of selection biases and the experimental variable: It is necessary to acknowledge that independent variables can interact with subjects that were selected from a population. For example, some factors (such as stress and multitasking) have different effects on memory in older than in younger subjects. In this case, the

outcome or findings from the experiment may not be generalized to a larger or different population.

Reactive effects of experimental arrangements: An experimental situation itself can affect the outcome, making it impossible to generalize. That is, the outcome can be a reaction to the specific experimental situation as opposed to the independent variable.

Multiple-treatment interference: If multiple-treatments should be applied to the same participant, the participant's performance would then not be valid because of the accumulated effects from those multiple treatments. For example, if you have learned sample material one way, it is hard to tell if later learning is the result of the new learning method presented second, or the result of the first method, or the combination of the two.

Why mention these effects in a book on how to run subjects? Why not just let these be mentioned in experimental design text or course? We mention them here because if you are new RA, you may not have had an experimental design class. And yet, many of these effects will be most visible to the person running the study. For example, if there is an event, such as an election, where you are running subjects, and you will be comparing the results with those from a different country where the PI is located and there is not an election, it is the RA that has the best chance of noticing that something unusual is happening that could pose a threat to the study's validity.

4.2 Risks to internal validity

There are other issues that investigators need to consider, such as participants' effects and experimenter effects. We will take these issues up in the following section.

4.2.1 Power: How many participants?

Human performance is noisy. Differences that appear could be due to a theoretical manipulation, or it could be due to chance. When piloting, you might start running and not have an endpoint in number of subjects in mind. This might also apply with informal controlled observation for interface and system development. With more formal studies, you will have to get approval (IRB in the US) for a set number of subjects. How do you choose that number?

There are two ways to approach how many participants to run. One way is through comparison to similar research and rules of thumb, and the other is through computations of statistical power. The heuristics are often used, and the power calculation assumes that you have an idea of what you are looking for, which you might not because you are still looking for it!

Each area will have its own heuristics for the number of subjects to run. The number to run is based on the hypothesis and the size of the effect for a given manipulation. In cognitive psychology near Rich Carlson, he suggests 20 subjects per condition. In human-robotic studies it appears to be between 20 and 40 (Bethel & Murphy, 2010). In (expert) heuristic interface evaluation, the number can be said to be 7 (Nielsen & Molich, 1990), but the range of user types is also important (Avraamides & Ritter, 2002). In physiological psychology, where people vary less, the number might be as low as 4 per condition. In areas with more subtle effect sizes, such as education, the numbers need to be larger.

The other way to determine the number of subjects to run is to do a power calculation based on the effect size you are looking for (Cohen, 1992). An effect size is how much does a change in the independent variable leads to a change in the dependent variable. The unit of measure used is the standard deviation in the data. An effect size of 1 is thus that the mean changes by a standard deviation. A standard deviation is about a grade in the traditional US grading scheme. Thus, an

effect size of 2 is a large size (comparable to being tutored individually), and an effect size of 0.1 is a smaller effect size.

This the intention behind statistical tests, to find out if the changes that we see arise from chance or are so extreme that they are unlikely to have arisen from chance. We now discuss the power of a statistical test, and how a test's power can influence its effectiveness. Calculating the test's power can help maximize the benefits of an experiment by helping you decide how many subjects to run. For instance, while relatively rare, running too many subjects can be wasteful when the effect size is known to be large.

Testing a hypothesis produces two outcomes: (a) one outcome rejects the null hypothesis ( ), while the other outcome (b) accepts the null hypothesis—that is accepting the alternative hypothesis (H_a). When investigators decide to either accept or reject the alternative hypothesis, they can make two types of errors, known as Type I and Type II errors. Table 4.1 describes these errors.

Table 4.1. Type I and II error in testing the null (H_0) and experimental (H_a) hypotheses.

	True State	
Decision Made	H_0 is true	H_a is true
Reject H_0	Type I error (report a result, but no effect)	Correct decision
Fail to reject H_0	Correct decision	Type II error (report no result, but there is an effect)

In fact, if the null hypothesis (H_o) is true, investigators should fail to reject the null hypothesis. When the null hypothesis is incorrectly rejected, Type I errors occur. The probability of making a Type I error is denoted by alpha, written α. On the other hand, if the alternative hypothesis (H_a) is true, in fact, investigators should accept the alternative hypothesis. When the alternative hypothesis is incorrectly rejected, Type II errors occur. The probability of making a Type II error is denoted by beta, written β. Experimenters will talk about Type I and Type II errors, so it's worth learning what they are.

The power of a test is defined as the probability of correctly rejecting the null hypothesis (H_o) when it is in fact true—this is denoted by $1-\beta$. In a practical sense, via the calculation of the power, investigators are able to make a statically supported argument that there is a significant difference when such a difference truly exists. Good sources of determining the size of a study *a priori* include: Cohen's work (in the further readings), explanations about study size and power in stats books (e.g., Howell, 2007, ch. 8), and programs that can be found online for free, such as G*Power3.

An important point to remember about statistical power is this: Failing to reject the null hypothesis is not the same as proving there is no effect of your independent variable. Knowing the statistical power of your experiment can help you ensure that if there is an effect, you will be able to find it.

Statistical tests such as ANOVA that involve null-hypothesis testing are standard in much behavioral research, but it may be useful to know that a number of researchers advocate

alternative approaches. One example is Bayesian analysis, which evaluates the strength of evidence for alternative hypotheses. Some researchers argue it is better to report mean results with some indication of the reliability of those means, such as the standard error. In many psychology journals, it has now become standard for editors to require researchers to report *effect sizes*, which are statistics measuring how big the effect of an independent variable is, relative to random variation in the data (Wilkinson, 1999). Another approach emphasizes graphic depiction of data with confidence intervals (e.g., Masson & Loftus, 2003). Other researchers argue that analyses should focus on just on the means but on other properties of data distributions. Some psychologists (Wagenmakers & Grünwald, 2006) have argued that Bayesian approaches should replace standard significance tests. In all cases, the recommendations for alternatives to null-hypothesis testing result from concerns about the nature of inference supported by findings of statistical significance. Elaborating on these alternatives is beyond the scope of this book, but it is useful to know that they are becoming increasingly prominent in behavioral research.

4.2.2 Experimenter effects

When two or more experimenters are running the same experiment, effects or biases from experimenters can exist. One experimenter may unconsciously be more encouraging or another more distracting in some way. Preventing possible experimenter effects is necessary for guaranteeing the validity of the experiment, both for the ability to repeat it and to generalize from it. Mitchell and Jolley (2012) note some reasonable causes for error that investigators should avoid: (a) the loose-protocol effect, (b) the failure-to-follow-protocol effect, and (c) the researcher-expectancy effect.

First, to avoid the loose-protocol effect, when you run the experiment and particularly when a study is run by different experimenters, it is necessary to write down the procedures in detail. The protocol document should allow other experimenters to run the experiment in exactly the same way, providing a standardized way to run the trials. Once you finish a draft of the protocol document, you should test it with practice participants. An example is included as an Appendix. Producing the final protocol document will require a few iterations of writing and testing the protocols with practice participants, revising the protocol in response to the first pilot trials.

The second cause of error, the failure-to-follow-protocol effect, results from an experimenter's failure to follow the experiment's protocols. There might be several reasons for not following the protocol—the reasons can include a lack of motivation to follow the protocol, or ignorance of the protocol, etc. Sometimes a failure to follow protocol can result from efforts to help the subjects. For example, one study found that the subjects behaved unexpectedly, in that they had fewer problems than were expected. Upon further investigation, it turned out that the student research assistants were breaking up the lessons into subparts to facilitate learning (VanLehn, 2007).

The third cause for error, the researcher-expectancy effect, arises from the influence of the experimenter's expectations upon his or her interactions with the participants. For instance, I might be biased (consciously or unconsciously) in how I run the experiment if I know I am testing my hypothesis. After all, I have a personal incentive to reject the null hypothesis in this case. Therefore, it is preferable when possible that the experimenters interacting with the subjects be unaware of the hypothesis being tested. When this happens, it is called a double-blind study; in this kind of study, the experimenter and the subject both do not know what treatment the subject receives. An example of a double-blind study would be when the RA and the subject both do not know which amount of caffeine a subject received, or to which condition the subject belongs.

Following consistent clearly written protocols in an unrushed manner is one way to avoid many of these errors. Please be patient and give the participants enough time to complete each procedure to the best of their ability.

4.2.3 Participant effects

Because personal characteristics and histories influence performance, it is important to try to methodically achieve a representative sample when selecting participants. Factors such as ethnicity, gender, age, experience, native language, or working memory capacity can all affect performance. Random assignment of subjects to conditions generally helps mitigate this effect. Random assignment, however, can go wrong (or be done incorrectly), or result in a suboptimal distribution. RAs often are the earliest, best, and often the only way to discover these problems.

4.2.4 Demand characteristics

Sometimes, internal validity is threatened by subjects' interpretation of the experimental situation. For example, a subject may think that he or she has figured out your hypothesis and deliberately attempts to be a "good subject" and provide the data he or she thinks you want. Conversely, some subjects may try to behave in way contrary to what they perceive as the hypothesis. For example, one of us, in conducting a study on causal reasoning, was surprised to hear a subject say "I'm wrecking your hypothesis—I'm behaving exactly the same way whether you look at me or not!" More subtly, subjects may perceive what you think is an innocuous manipulation as an attempt to induce stress, or a set of questions about an interface as an effort to measure personality. Very often, subjects recruited from university subject pools have read about research using deception, and assume that all experiments involve some kind of deception. Even worse, demand characteristics can influence behavior even when the subject is not aware of their influence.

The properties of experimental situations that lead subjects to try to behave in certain ways have been labeled *demand characteristics*—that is, characteristics of the situation that seem to demand certain kinds of behavior or the adoption of particular roles. The term was introduced by Martin Orne in the 1960s, and an encyclopedia entry he co-authored provides a brief summary of the concept (Orne & Whitehouse, 2000).

Detailing the wide variety of possible demand characteristics is beyond the scope of this book. However, we can offer some general advice. First, being aware of possible demand characteristics may help you to avoid them. It is useful, for example, to ask a few pilot subjects who are naïve to your hypotheses what they thought your experiment was about. Second, clear instructions with as much openness as possible about the purpose of the experiment will help avoid misinterpretations of the task. Sometimes it is even useful to explicitly say that the goal of the experiment is not to assess personal characteristics such as personality or intelligence (assuming, of course, that is true). Third, greeting subjects in a friendly and natural way may help avoid suspicions of deception.

4.2.4 Randomization and counterbalancing

Randomization describes the process of randomly determining both the allocation of the experimental material and the order in which individual trials are to be performed (Montgomery, 2001). Random assignment refers to assigning subjects to experimental conditions so that individual differences are not correlated with the independent variables (e.g., sex, order of arrival). In all of these cases, the basic idea of randomization is to control for factors that might affect your dependent variables, but are neither explicitly controlled by setting the levels of your independent variables nor of direct interest to your research question. Thus, the effect of

individual differences, particular stimuli and order are "averaged out" across the conditions of your experiment, helping to maintain internal validity. Of course, the larger your sample of subjects, your sample of experimental materials, or the number of alternative orders, the more effective this averaging-out process is. Statistical methods assume that the observations are independently distributed random variables. Proper randomization of the experiment helps in making sure that this assumption is at least approximately correct, and allows us to conduct standard psychological tests.

Failing to randomly assign subjects to conditions can cause a number of problems. For example, it might be convenient to run one experimental condition in the morning, and another in the afternoon. However, the subjects who sign up to participate in experiments in the morning are likely to be systematically different than those who sign up for afternoon sessions. Researchers who use university subject pools are familiar with the time-of-semester effect: subjects who sign up for studies earlier in the semester are often more motivated and conscientious than those who sign up later in the semester.

Random sampling, a related term, is a method for selecting the entire sample group. Ray (2003) notes that one way to achieve external validity is to have the participants in the experiment constitute a representative sample of the entire population. In fact, it is very hard to accomplish true random sampling. However, it is useful to plan recruiting so as to minimize such biases, as discussed in Section 4.3.2.

Montgomery (2001) notes that in some situations it is difficult to achieve true randomization because of a hard-to-change variable (e.g., the subject's gender). Sometimes, it may be useful to use what is known as constrained randomization. For example, you might randomly assign subjects to experimental conditions with the constraint that an equal proportion of male and female subjects are assigned to each condition. Similarly, if you have two conditions that are manipulated within subjects (that is, each subject experiences both conditions), rather than randomization you might assign equal numbers of subjects (randomly, of course) to the two possible orders. This strategy is known as *counterbalancing*.

Practically, there are several ways to randomly assign subjects. One easy way is to create a way to randomly assign subjects to condition. For two conditions, it can be a coin, for 3, and 6 conditions, a die can be rolled. For more conditions, you can use a random number generator or a deck of playing cards or some note cards made for the purpose of the study. If you have 30 subjects, roll the die 30 times, or shuffle the cards and deal out 30 cards (perhaps from a smaller deck). The order of the cards, dice, coins gives you the order of assignment. You should check that the balance is correct, that is, that you have equal numbers of each conditions. You may also use a table of random numbers (found in many statistical textbooks) or computer software that generates random numbers (most spreadsheets can do this), or you can randomize an array of numbers. Random-assignment features may also be included in software packages designed for behavioral research.

Remember that you are better served by doing balanced assignment, that is, equal assignment to each group. A pure coin flip will not ensure this because in a series of 10 trials there will not always be 50% heads and tails, and you are better served by doing assignment without replacement. So, creating a set of assignments and then randomly ordering them will work more naturally and efficiently.

Randomization and counterbalancing apply not just to the assignment of subjects, but to the arrangement of stimuli and experimental conditions. For example, if you are conducting a memory study in which subjects are learn a list of words or other materials, there might be effects of the order in which the material is presented. By presenting the items in a new random order for each subject, any effects of order will be balanced over subjects. Similarly, when an

independent variable is manipulated within subjects, you may want to assign some stimuli to one level of a variable and some to another. Reversing the assignment for half of the subjects is an example of counterbalancing. There are many possible randomization and counterbalancing schemes, and choosing one will depend on the details of your experiment. In general, randomization is effective when there are many opportunities for different random orders or arrangements. When there are only a few such opportunities, counterbalancing is preferred because it guarantees that the factors you are counterbalancing, such as the assignment of stimuli to conditions, are equally distributed over levels of your independent variables (that is, occur equally often for each level).

4.2.5 Abandoning the task

During the experiment, subjects' minds may wander from the task. Few will stop using a computer-based task if you are in the room (so, if this problem occurs, stay in the room!). Some will pay less attention, and one way to avoid this is to have shorter experiments. It is also important to strive to run a crisp experiment where your professional bearing and expectations indicate the necessary sense of gravity leading them to try to do well.

In experiments using verbal protocols, the subjects may stop talking or talk about other topics. You should neither let them sit without talking nor let them talk about non-task related things. In the first case, you need to ask them to "keep talking" (Ericsson & Simon, 1993, Appendix). In the second case, if they wander off from reporting their working memory onto other topics, you may have to ask them to focus on the task. Asking them to do these things is highly appropriate, and if you do not you will hurt the experiment. You might be more comfortable if you practice this with both a helpful (compliant) and unhelpful (incompliant) friend as pilot subjects. It is also very appropriate and helpful to put the conditions for such prompts into your script.

Finally, if the subject does wish to completely abandon the task, you need to let them do that. In nearly all study protocols, they receive full compensation if they start. Withdrawing from a study of the type discussed in this book is rare, but it needs to be accommodated gracefully and graciously. Some subjects will be taking advantage of the situation and some who abandon the task will have become uncomfortable in some way, and you cannot really tell them apart. In either case you have to treat them both kindly. Persuading a subject to continue when he or she wants to withdraw may be seen as inappropriate coercion, which raises ethical problems. If it helps, keep in mind that getting a reluctant subject to stay may encourage erroneous data.

4.3 Risks to external validity

It is possible, even common, to run an experiment with excellent internal validity, only to find that your conclusions do not apply in other situations where you think they should. For example, you might conclude from your experiment that factor X influences behavior Y…and find that, in fact, the conclusion is inapplicable outside of a lab environment because it only applies to students at your university, or who are from that high school program. This lack of ability to generalize the results is a failure of external validity, or generalizability. We discuss some of the common problems that cause such failures.

4.3.1 Task fidelity

Most experiments use tasks that are meant to capture some important aspect of behavior in real-world settings: for example, how feedback affects learning, how display features affect the guidance of visual attention, or how aspects of an interface influence task performance. Usually, though, the experimental task is a simplified version of the real-world situation. Simplifying the task environment is a good and often a necessary step from the perspective of internal validity—it

is easier to establish effective experimental control when you use a simplified task. However, if the task fails to capture the critical features of the real-world situation, your results may not apply to that situation. The degree to which it succeeds in doing so is known as *task fidelity*.

Sometimes, the issue of task fidelity is addressed by using an experimental task that is almost indistinguishable from the relevant real-world situation. For example, recent research on driver distraction often uses very high-fidelity driving simulators that use actual automobile dashboards and controls, with high-definition displays that update in response to the simulated motion of the car and the actions of the driver. Such simulators allow both excellent experimental control and an opportunity for excellent external validity. Other research makes use of virtual-reality setups that allow similar complexity and fidelity. However, such simulators and virtual-reality setups are expensive and impractical for most research, and it is not always clear what features are needed.

What do you do if you can't arrange high-fidelity simulations of real-world environments? The best answer is to make sure you have *psychological fidelity*—that is, that your experimental task accurately captures what is psychologically and behaviorally relevant about the real-world situation. This involves several aspects of the experimental task and its relation to the real-world situation. First, you should consider whether the information available to subjects, and the behavior requested of them is *similar* to the real-world situation—do these resemble the real situation in terms of the perceptual, cognitive, and motor aspects of the behavior? Second, you should consider whether the experimental situation is *representative* of the real situation—for example, do the cues available predict other cues or outcomes in the same way (with the same probability, or subject to the same contextual influences) as in the real situation? This may require careful thinking about the frequency with which subjects encounter particular stimuli, for example. Psychological fidelity thus involves both resemblance and structure (for further discussion, see Dhami & Hertwig, 2004; Kirlik, 2010; Smallman & St. John, 2005). It may be easiest to understand the issue of psychological fidelity by considering some examples in which generalization has or has not been successful.

Early in the history of research on memory, Ebbinghaus (1885/1964) decided that he could best achieve experimental control by using nonsense syllables (syllables such as GAX that have no meaning) to avoid the influence of prior learning. However, in real-world memory situations, people rely on associations to prior knowledge as a basis for remembering. Thus many of Ebbinghaus's results, while they can be repeated, are difficult to generalize to the real world. For example, he provided an elegant description of the relation between repetition and memory, but in the real world, some things are memorable after a single repetition while others are much harder to learn.

Studies of learning and training provide many examples of both successful and unsuccessful generalization. In these studies, the research question often focuses on whether learning in a simplified or simulated training environment can be transferred to a real task environment, and whether the variables that affect learning in the training environment predict performance in the real task environment. For example, Cassavaugh and Kramer (2009) reported that the effects of training in relatively simple tasks generalized to driving by older adults in a high-fidelity driving simulator. On the other hand, Lintern, Sheppard, Parker, Yates, and Nolan (1989) found in a study of simulator training for military flight that the simulators that produced the best performance in training were not the ones that resulted in the best performance in actual flights. These results suggest that careful analysis of the perceptual and cognitive components of the task (of driving and flying) need to capture the relevant similarities to be generalizable.

The point is that evaluating the external validity of your experimental task is not a simple question, and must be answered not by a subjective judgment of how similar your task is to the

real-world situation of interest, but by a systematic consideration of what aspects of the task are important. Proctor and Dutta (1995, Chapter 9) provide a useful introduction to this issue in the context of training research.

4.3.2 Representativeness of your sample

We mentioned earlier the value of recruiting a broad sample of subjects. The field of psychology has often been criticized for conducting research primarily with college students in Western cultures. While there is a good argument that restricting research to this kind of sample is fine for studying very basic processes that are not expected to differ from person to person, that argument breaks down when we consider many research questions that are relevant to real life. For example, older individuals often employ different strategies for prospective memory (remembering to do things) than do college-age subjects. Icons and instructions that are easy for American college students to interpret may be completely opaque to people living in other cultures.

One way to get a more representative sample is not to describe many details of your study. If you note that it is 'math puzzle fun', you will get subjects who are interested in math. If you note a study about cognition alone you will get less self-selection by the potential subjects.

Listing all of the ways in which a restricted sample of subjects can make it hard to generalize experimental results could fill a book (e.g., Jonassen & Grabowski, 1993). The important point is that you should think about the situations to which you want to generalize, and ask yourself how your sample might differ from the general population in those situations. The best thing, of course, is to recruit subjects who are representative of the population of interest.

4.4 Avoiding risks in the multilingual fonts study

As noted previously, internal validity refers to whether we can assert with any confidence that changes in the independent variable reliably lead to changes in the dependent variable or variables. To establish internal validity, we must show at a minimum that cause precedes effect (temporal precedence), that cause and effect are related (covariance), and that no plausible alternative hypothesis exists that better explains the correlation found between the variables (nonspuriousness). Threats to internal validity are any factor or combination of factors that introduces ambiguity as to the nature of the relationship being studied. These threats can include but are not limited to: confounding variables, selection bias, historical or situational factors, maturation, repeated testing, and experimenter biases.

If we examine Ying's and Edward's study, we find that threats to internal validity can emerge from both serious contextual issues that require explicit steps be taken during the design, approval, and recruitment stages of the experiment, as well as avoiding seemingly innocuous oversights that nevertheless can jeopardize the experiment's internal validity. Counteracting these threats requires the vigilance, cooperation, and sometimes creativity from the whole team. We will discuss both sets of problems within the context of this study here and in Chapter 5.

At the onset, Ying and Edward faced a serious problem, achieving a representative sample size. After completing a power analysis (Cohen, 1988, 1992), Ying found that she needed at least 30 participants per alphabet group (Arabic and Hangul), a minimum of 60 participants. Previous studies examining matrix formats for non-roman alphabets have primary occurred outside of the US. Finding participants with fluency in the alphabet of interest posed a significant challenge. This challenge was further complicated by the external validity concerns of key stakeholders within the OLPC, whether the results would be generalizable outside of the experimental setting. To satisfy these stakeholders, a condition for the study was to test all participants using interfaces featuring the screen resolution and size found on OLPC machines, to ensure the matrix formats

would be evaluated under conditions similar to that of the children in the program. While keeping the screen size and resolution constant across experimental conditions was necessary for internal validity, usually HCI studies feature larger (desktop) screens with better resolution. In either case, the need to control for both factors made an online study impractical.

Using fliers and the departmental newsletter enabled Ying and Edward to find enough participants for the pilot study, 15. These methods alone, however, were insufficient to get the 60 participants necessary for the study. Ultimately, Ying and Edward had to contact student groups associated with these populations. Consequently, while the participants all had sufficient fluency in English to complete the study, there were participants—generally friends and family of graduate students- who required additional instructions and handouts in their own languages, as well as maps to the lab. In addition, scheduling participants required flexibility and a willingness to accommodate parental obligations of young parents. But, by planning and getting more resources (including time) than had originally been budgeted, the study could be completed.

4.5 Avoiding risks in the HRI study

In preparing his study and recruiting subjects, Bob particularly needs to worry about risks to external validity: do the results have impact for his users, his robots, and the tasks and situation in which they will be frequently used? Bob's company will be interested in the generalizability of his results, and not simply whether the results speak to the particular situation he was able to study. So, he should take care that the subjects he uses are similar to the robot's potential users. If the robot is for the elderly or for children, he should have elderly or children users. He should not have the engineers who already know how to use it because they helped build it (although some of them will think that they are like the target users or think that they can "pretend" to be like them). He should also be careful if he includes friends or neighbors of the engineers among his participants. These people may have spoken with the engineers or worked with them, and might know too much to accurately represent the user population.

Bob should also take some care that the tasks and the environment are operationally similar to the tasks required of the robot and its operators in their environment. If the robots are for engineers in R&D firms, then, he is set because the robots are in their natural setting. If, on the other hand, the robots are for disaster relief workers, he will need a situation and users similar to the situations the robot and its operators will have to face, for example, a pile of rubble and fireman (not undergraduate or graduate students) to help test the robots, see Murphy's work, for example (Murphy, Blitch, & Casper, 2002).

4.6 Conclusion

We have discussed here some of the major threats to internal and external validity. There are too many possible threats to address them all, but being aware of the types of threats can help you design a better experiment. Perhaps more important, thinking about the threats discussed here can help make you aware of possible limitations of your experiment, and let you recognize other threats to validity we have not discussed.

4.7 Further readings

Here is a list of further reading materials concerning this chapter.

- Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.
  Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*. 98-101.

Cohen has raised the issue of the importance of statistical power analysis since the 1960's. He originated the statistical measure of power, that is, of measuring the effect of a manipulation in terms of the natural variation in the measurements, effect sizes. The two articles above will help you be aware of this issue and avoid Type I and II errors in your research experiments.

- Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson.

Howell's book provides a useful summary of how to apply power written for those learning statistics. Other introductory statistics books will have similar treatments. They are useful introductions to this process.

- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Cambridge, MA: MIT Press.

Ericsson and Simon in this book explain the theory of how verbal protocols can be used, and in what ways they are valid, and when they are invalid.

4.8 Questions

Summary questions

1. Answer the following questions.
   (a) What is "the experimenter effect"?
   (b) What is "randomization"?
   (c) What is "fraud"?
   (d) What is "generalizability"?
   (e) What is "effect size"?

2. List 12 factors that can endanger the validity of your research with human subjects (i.e., internal and external validity issues).

3. Explain Type I and Type II errors in testing a research hypothesis.

Thought questions

1. Recall the question number 2 in Thought Questions, Chapter 1 (the operational definitions of the research variables). Suppose that you will conduct a research study with these variables. Discuss whether there are risks that might endanger the validity of your research. Discuss how you plan to mitigate the risks.

2. If you run a study using university freshmen, explain what this will mean for your results. If you run a study using people recruited from a newspaper ad, explain what this will mean for your results.

5 Running a Research Session

This chapter provides practical information on what to do when you run your experiments. We assume that you have developed your initial experimental design and are now ready to run a pilot study. This chapter is thus about interacting with subjects and the context in which you do that.

5.1 Setting up the space for your study

The environment you provide for your subjects is important in making sure your data is of high quality. Typically, setting up the space for your experiment will seem straightforward—often, subjects will simply sit at a computer performing the experimental task. However, giving some thought to setting up the space in advance can help. For example, if possible, you should provide an adjustable-height chair if subjects are sitting at a computer. Avoiding screen glare from overhead lights can be important—it may be helpful to have an incandescent table lamp to use instead of bright fluorescent ceiling fixtures. Allow for the possibility that some of your subjects may be left-handed—we have seen experimental setups that were very awkward for left-handers to use. In general, try to take the perspective of your subjects and make the setup as comfortable as possible for them.

In setting up the space, it is also important to consider possible distractions. For example, if your experimental space is next to an office, or opens on a busy hallway, consider the possibility that loud conversations nearby may distract your subjects. The ideal setup for running individual subjects is a sound-isolated chamber or room, but that is not always practical. A simple sign that reads "Experiment in Progress—Quiet Please" can help a great deal. If you must collect data in a room that is also used for other purposes, such a sign can also help avoid accidental intrusions by others who may not realize that an experiment is in progress. (Also, take the sign down after the study.) It is also best to avoid "attractive nuisances" in the experimental space—things that are inviting to inspect. For example, one of us collected data in a room that had a shelf full of toys and puzzles used in another study—until we found a subject playing with a puzzle rather than performing the experimental task!

Often, subjects may have to wait after arriving at your study, perhaps as other subjects finish. Though, of course, you should try to minimize waiting time—unlike a doctor's office or drivers license center, your subjects don't *have* to be there—it is important to provide a comfortable place to wait. If the only waiting area available is a hallway, try to at least to place chairs in an appropriate location with a sign that says "Please wait here for TitleOfExperiment experiment."

Figures 5-1 and 5-2 show two spaces used for running subjects in a psychology department. Figure 5-1 shows a small storage space used a single-subject data collection station. A table lamp is used to avoid glare from overhead fluorescent lights, and the room is free of distractions. The room is on a quiet, rarely used hallway, so this space provides good isolation. A nearby workroom serves as a reception and waiting area, as well as office space for research assistants.

Figure 5-2 shows a large office used to house multiple data-collection stations. Office dividers separate the stations and provide some visual isolation, while allowing a single experimenter to instruct and monitor several subjects simultaneously. In such setups, subjects are sometimes asked to wear headphones playing white noise to provide additional isolation. In this space, subjects wait for their sessions in the hallway, requiring a sign asking for quiet.

Figure 5-1. A storage space used as a single-subject data collection station.

Figure 5-2. An office space used to house multiple data-collection stations.

5.2 Dress code for Experimenters

The goal of a dress code is to convey a serious atmosphere and to encourage respect and cooperation from your subjects. You should consider the impression you wish to make and will make when running your experiment. This consideration should include you with to position yourself (as to command respect while making the participants comfortable enough to perform the task), the type of experiment, and the type of participants in the experiment.

In most cases, we recommend wearing a semi-professional outfit, such as a dress shirt with dress slacks, when running experiments. This helps you look professional and prepared but not intimidating. Semi-professional dress helps convey the experiment's importance while not overwhelming the participant. However, appropriate dress may vary depending on your subject

population. If you are a college student interacting with college-student subjects, it may be best to dress like a college student—but think of a college student who wants to make a good impression on a professor, not a college student hanging out in the evening. It is certainly best to avoid things like t-shirts with slogans some might find offensive, low-cut blouses, very short shorts or skirts, or flip-flops. If you are working with non-student adult subjects, "business casual" is a better choice of dress. If your subjects are expert professionals, you should dress in a way that would fit in their workplace.

5.3 Before subjects arrive

Your interaction with the subjects you've recruited begins before they arrive to participate. It is important to be clear about when the study is. It is wise to remind subjects by phone or email the day before a study is scheduled, if they have been scheduled farther in advance, and to repeat the time, place, and directions in the reminder. If there is a time window beyond which you cannot begin the study—for example, you might want to exclude from a group study anyone who arrives more than 5 minutes late—make sure this is clear as well.

As you schedule the times to run you should take advice about when to schedule times. It is usually appropriate to schedule times during normal business hours (which in a university lab may be 10 am to 6 pm). If you are running subjects outside of these normal hours you should have a discussion with the principal investigator about safety for you and for the subjects (how to reach the PI, for example). You should also consider practical issues such as whether the building will be locked after normal business hours or on weekends. If your subjects are traveling some distance to be in your experiments, do parking restrictions or bus schedules change after hours or on weekends?

Make sure that your subjects have clear directions to the location of your study. On a college campus, it may be important to provide directions and identify nearby landmarks. If subjects are driving to the location of your study, make sure you provide clear instructions on where to park and whether they are expected to pay for parking. Make sure the door to the building is unlocked, or have someone meet subjects at the door—one of us knows of an experiment in which several subjects were not run and hours of data collection were lost because the experimenter didn't realize that the campus building would be locked after 5 p.m, and the subjects were literally lost.

You should also provide clear directions to the specific room in which the study is held. One of us works in a psychology department, and not uncommonly sees research subjects wandering the halls looking for the room their experiment is in. It also helpful to clearly mark the place where the experiment will be (or the place where subjects should wait)—a simple sign that says "Skill Acquisition Experiment here" may save a lot of confusion in a building where every hallway and doorway looks pretty much alike and there are multiple experiments. If subjects must pass a receptionist to find your study, make sure the receptionist knows where the study is and who is running it—many people will stop to ask even if they think they know where they're going.

Making it as easy as possible for subjects to find your study and to arrive in a timely way is important for ensuring that they arrive ready to participate, with minimal anxiety. This helps in establishing the cooperative relationship with your subjects that will yield the best results for your experiment.

5.4 Welcome

As the experimenter, you are taking on a role similar to that of a host, thus, it is appropriate to welcome participants to the study. Where it is appropriate, you might provide them materials to read if they have to wait, and to answer questions they have before the study begins. It is also

very appropriate to confirm their names (for class credit), and to confirm for them that they are in the right place and at the right time. If the experimental protocol permits it, you might also indicate how long the study will take. This helps set the stage for the study itself.

The first event after welcoming subjects is typically the informed consent procedure. It is important to take this seriously—while it will become routine to you, it is likely not to your subjects. Rather than simply handing a subject the consent document and saying "you have to sign this before we can start," take the time to explain the major points, and to provide an opportunity for questions. Many will have no questions, glance quickly at the document, and sign it. Nevertheless, your approach every time should be one that allows the subject an opportunity to understand and to think about what they are agreeing to.

5.5 Setting up and using a script

Your research study will likely have a script of how to run the session. If it does not, it should, and it will help you run each subject in a confident and consistent manner. The script will often start with how to setup the apparatus. Before the subject's arrival, the experimenter needs to set up the apparatus and should be ready to welcome the subject. Incorrect or inconsistently applied procedures of the apparatus setup can sometimes cause inconsistencies in running the experiment (e.g., omission of a step). Consequently, the script that appropriately represents required procedures plays an important role in conducting a successful experimental study. Appendix D provides an example study script.

The setup should include making sure that all materials that are used are available (e.g., forms, at least one back up copy), and that the apparatus is working. If batteries are used in any of the apparatus (e.g., a laser pointer, a VCR remote), spare batteries should be to hand.

5.7 Talking with subjects

When you first welcome the subjects to your study and the study area, you might feel uncomfortable. After you have run a few sessions, this discomfort will go away. In a simple study, you can be quite natural, as there is nothing to 'give-away'. In more complex studies, you will be busy setting up the apparatus, and this tends to make things easier. It is important, however, to realize that talking with subjects before they begin the experiment plays an important role in getting good data. Often, subjects come to the lab feeling nervous, with little or no experience in participating in research and, perhaps, misconceptions about the nature of behavioral research. For example, it is not unusual for students participating in university subject pools to believe that all experiments involve deception, or that all researchers are surreptitiously evaluating their personalities or possible mental disorders. Interacting in a natural, cordial way, and explaining clearly what your subjects will be asked to do can go a long way toward alleviating the subjects' anxiety and ensuring that they do their best to comply with the instructions and complete the experimental task. In our experience, it is all too easy for experimenters to interact with subjects in a rote manner that increases rather than alleviates their anxiety. Remember that although you may have repeated the experimental protocol dozens of times, it is the first time for each subject!

In nearly all cases, abstaining from extraneous comment on the study is an important and useful practice that makes all parties concerned more comfortable. Many experimental protocols require not giving the subject feedback during the study. In these cases, your notes will probably indicate that you tell the participants at the beginning of the session that you are not allowed to provide them feedback on their performance. Generally, the debriefing can handle most questions, but if you are not sure how to answer a question, either find and ask the investigator, or, take contact details from the subject and tell them you will get them an answer. And then, do it! This also

means that when you are running subjects for the first couple of times that someone who can answer your questions should be available.

In social psychology studies or where deception is involved, you will be briefed by the investigator and will practice beforehand. In this area, practice and taking advice from the lead researcher is particularly important.

Be culturally sensitive and respectful to the participants. Consult with the lead investigator if you have general questions concerning lab etiquette, or specific questions related to the study.

There are a few things that seem too obvious to mention, but experience tells us that we should bring them up. Don't ask a subject for his or her phone number, no matter how attractive you find them! The experiment is not an appropriate context to try to initiate a romantic relationship. Don't complain about how hard it is to work in the lab, or how difficult you found your last subject. Don't tell a subject that his or her session is the last session of your workday, so you hope the session is over quickly. And so on. It might seem that nobody with common sense would do any of these things, but we've seen them all happen.

5.6 Piloting

As mentioned earlier, conducting a pilot study based on the script of the research study is important. Piloting can help you determine whether your experimental design will successfully produce scientifically plausible answers to your inquiries. If any revision to the study is necessary, it is far better to find it and correct it before running multiple subjects, particularly when access to subjects is limited. It is, therefore, helpful to think of designing experiments as an iterative process characterized by a cycle of design, testing, and redesign as noted in Figure 1-1. In addition, you are likely to find that this process works in parallel with other experiments, and may be informed by them (e.g., lessons learned from ongoing related lab work may influence your thinking).

Thus, we highly recommend that you use pilot studies to test your written protocols (e.g., instructions for experimenters). The pilot phase provides experimenters the opportunity to test the written protocols with practice participants, and is important for ironing out misunderstandings, discovering problematic features of the testing equipment, and identifying other conditions that might influence the participants. Revisions are a normal part of the process; please do not hesitate to revise your protocols. This will save time later. There is also an art to knowing when not to change the protocol. Your principal investigator can help judge this!

The major reason for returning to the topic of piloting here is that the pilot study provides an opportunity to think through the issues raised here—the setup of the experimental space, interacting with subjects before, during, and at the conclusion of the experiment, and so on. Especially for an inexperienced experimenter, pilot testing provides an opportunity to practice all of these things. In some cases, it may be effective to begin pilot testing with role-playing—one member of the research team plays the role of the subject, while another plays the role of experimenter.

You will often start piloting with other experimenters, and then move to officemates and people down the hall. One researcher we know gets IRB approval early and switches to subjects that could be kept using them as pilot subjects, and when the process is smooth declares them as keepers. This is expensive, but for complicated studies is probably necessary because your lab mates know too much to be useful pilot subjects. It is important to keep in mind that once you involve actual subjects whose data you may keep, or who are recruited from a subject pool, all of the issues concerning IRB approval discussed earlier come into play.

It is also important when piloting to test your data gathering and analyses steps. We have wasted significant amounts of resources when the apparatus did not measure what we thought it did, and we know of numerous studies where the format of the study software output did not load easily and directly into analysis software, or did not record the information that was later found to be needed. So, as an important part of piloting, take some of the pilot data and test analyzing it to see that the data is recorded cleanly and correctly, that it loads into later analysis tools, and that the results you want to examine can be found in the recordings you have. You can also see if your manipulations are leading to changes in behavior.

5.5 Missing subjects

In every study, there are two key parties—the experimenter and the subject or subjects (when running groups). Inevitably, you will encounter a situation where a participant does not show up despite having an appointment. While participants should notify you in advance if they are going to be absent, keep in mind that missed appointments do happen, and plan around this eventuality. Participants are volunteers (even when you consider compensation). Therefore, it is appropriate to be gracious about their absence. Where possible, we recommend offering to reschedule once. However, when there are repeated absences, it is often not worth rescheduling. Bethel and Murphy (2010) estimate that approximately 20% of subjects will fail to arrive. This seems slightly high to us; for example, in the Psychology Department subject pool at our university, the no-show rate is typically 5-7%. In any case, the lesson is that you will have to schedule more subjects than your target to reach your target number of subjects, particularly for repeated session studies, or studies with groups.

In some cases you as an experimenter may need to cancel an experiment. As an experimenter, it is unacceptable to simply not show up for an experiment. When you really have to cancel the experiment, you should do it in advance. Furthermore. as the experimenter, you have the responsibility to cancel the experiment by directly contacting the participants.

Note that in some cases, there will be specific rules about these issues—for example, the policies of your subject pool may require 24 hours notice to cancel a study, or have criteria for when absence is excused or unexcused. It is important to know and follow these rules.

5.8 Debriefing

The APA's ethical principles offer a general outline of debriefing procedures. For many experiments, the lead researcher may provide additional guidance. Investigators should ensure that participants acquire appropriate information about the experiment—such as the nature, results, and conclusions of the research. If participants are misinformed on any of these points, investigators must take time to correct these misunderstandings. Also, if any procedures are found to harm a participant, the research team must take reasonable steps to report and to alleviate that harm.

The experiment's procedures may cause participants to feel uncomfortable or be alarmed. After the experiment is finished, investigators or experimenters should listen to the participants' concerns and try to address these problems. Mitchell and Jolley (2012) provide reasonable steps to follow when you need to debrief:

- Correct any misconceptions that participants may have.
- Give a summary of the study without using technical terms and jargon.
- Provide participants an opportunity to ask any questions that they might have.
- Express thankfulness to the participant.

When you have a study that can be perceived as being deceptive or when the study is a double-blind study, you should seek advice about how to debrief the participants. If deception is a procedural component, you will most likely have to explain this to the subjects, and ask that they not discuss the study until all of the subjects have been run (the study's completion date). Requesting the participants to refrain from discussing the study will help keep potential subjects from becoming too informed.

To review, double-blind studies prescribe that neither the subject nor the experimenter knows which treatment the subject has received. For example, the amount of caffeine any single participant has ingested in a caffeine study with multiple possible doses. In these cases, you will have to explain the procedures of the study, as well as provide a general rational for double-blind trials. Otherwise, participants may balk at being given a treatment in a sealed envelope, or by a person who is not the experimenter. Furthermore, events such as the Tuskegee experiment (see Chapter 3) underscore why procedural transparency is so essential[12].

Reviewing your plans for debriefing will be part of obtaining approval for your experiment from the IRB or ethics panel. Sometimes, there are local rules about debriefing—for example, a university subject pool may require an educational debriefing for every study, even when the IRB does not. In an educational debriefing, you may want to briefly describe the design of the study and the theoretical question it addresses, using keywords that allow the subject to see connections between participating in your study and what they are learning in their psychology class. You may be required to provide a written debriefing, or to have your debriefing approved by the administrator of your subject pool.

As with the informed consent procedure, you may find that some, even most, subjects are uninterested in the debriefing. Also, debriefing will become routine to you as you run more subjects. It is important not to let these things lead you to approach debriefing in a perfunctory way that conveys to all subjects that you do not consider it important. If only one subject appears interested, that is reason enough to take debriefing seriously.

5.9 Payments and wrap-up

At the end of the session, you should be sure to compensate the subject as specified. Compensation can include monetary payment, credit towards a class, or nothing. If you are paying them monetarily, check with your supervisor, as there are nearly always detailed instructions for how to process such payments. In any case, you should make sure that they receive their compensation; you receive any required documentation such as receipts; and that you thank each participant for their assistance. Without them, after all, you cannot run the study.

At the end of the wrap-up, you should set up for the next subject. Make sure that copies of forms are to hand, and that if you have used such things as spare batteries you get some fresh batteries.

5.10 Simulated subjects

You may find yourself running simulated subjects. User models and simulations are increasingly used, both as standalone objects, but sometimes as part of a study to provide a social context. For example, to model a social situation you might have two intelligent agents act as confederates in a resource allocation game (Nerb, Spada, & Ernst, 1997). These agents provide a known social context in that their behavior is known and can be repeated, either exactly or according to a proscribed set of knowledge.

[12] The abuses associated with these studies led to the Belmont Report and the modern IRB process as a means of mitigating future risks to experimental participants.

When you run simulations as subjects, you should keep good notes. There are often differences between the various versions of any simulation, and this should be noted. Simulations will also produce logs, and these logs should be stored as securely and as accurately as subject logs. There may be more of them, so annotating them is very prudent.

If you create simulations, you should keep a copy of the simulation with the logs as a repeatable record of the results. You should keep enough runs that your predictions are stable (Ritter, Schoelles, Quigley, & Klein, 2011), and then not modify those files of model and runs but only modify copies of them.

Obviously, many of the issues discussed in this chapter do not apply to simulated subjects—no one, to our knowledge, has ever proposed that a simulated subject should be debriefed! Nevertheless, the importance of a clear protocol for your experiment is unchanged.

5.11 Problems and how to deal with them

For cognitive psychology and HCI studies, most studies run smoothly. However, if you run experiments long enough, you will encounter problems—software crashes, apparatus breaks, power goes out, and so on. Sometimes, too, there are more person-oriented problems—difficult subjects or problems that involve psychological or physical risks to the subject. Ideally, the research team will have discussed potential problems in advance, and developed plans for handling them. It is the nature of problems, though, that they are sometimes unanticipated.

The most common problems are minor—software or equipment failures, problems with materials, and so on. In responding to such problems, the most important things to remember are (a) remain calm—it's only an experiment, and (b) try to resolve the problem in a way that does not cause difficulties for your subject. For example, computer problems are often solved by rebooting the computer—but if this happens 30 minutes into a one-hour session, and you would have to start over at the beginning, it is not reasonable to expect the subject to extend his or her appointment by half an hour. Often, the best thing to do is to apologize, give the subject the compensation they were promised (after all, they made the effort to attend and the problem is not their fault. It is appropriate to be generous in these circumstances.), make a note in the lab notebook, and try to fix things before the next subject appears.

It can be harder to deal with problems caused by difficult subjects. Sometimes, a subject may say, "This is too boring, I can't do this…", or simply fail to follow instructions. Arguing with these subjects is both a waste of your time and unethical. As noted in Chapter 3, a basic implication of the voluntary participation is that a subject has the right to withdraw from a study at any time, for any reason, without penalty. Depending on the situation, it may be worthwhile to make one attempt to encourage cooperation—for example, saying "I know it is repetitive, but that's what we have to do to study this question"—but don't push it. A difficult subject is unlikely to provide useful data, anyway, and the best thing is to end the session as gracefully as you can, note what went on, and discuss the events with the PI.

You can also encounter unexpected situations in which a participant is exposed to some risk of harm. For example, occasionally a subject may react badly to an experimental manipulation such as a mood induction or the ingestion of caffeine or sugar. It is possible, though extremely rare, for apparatus to fail in ways that pose physical risks (for example, if an electrical device malfunctions). And very rarely, an emergency situation not related to your experimental procedure can occur—for example, we know of instances in which subjects have fainted or had seizures while participating in experiments, and fire alarms can go off. Investigators must be committed to resolving these problems ethically, recognizing that the well-being of the participants supersedes the value of the study. If an emergency situation does arise, it is

important that the experiment remain calm and in control. If necessary, call for help. If the problem is related to the experimental procedure, it may be wise—or necessary—to cancel upcoming sessions until the research team has discussed ways to avoid such problems in the future.

It is important to bring problems to the attention of the lead researcher or principal investigator. In the event of problems that result in risk or actual harm to subjects, it is important to consult the relevant unit responsible for supervising research, such as the IRB. These problems are called "adverse events" and must be reported to the IRB.

5.12 Chance for Insights

Gathering data can be tedious, but it can also be very useful. The process of interacting with subjects and collecting data gives you a chance to observe aspects of behavior that are not usually recorded, such as the subjects' affect, their posture, and their emotional responses to the task. These observations that go beyond your formal data collection can provide useful insights into the behavior of interest. Talking informally with subjects after they have finished the experiment can also provide insights.

Obtaining these kinds of insights and the intuition that follows from these experiences is important for everyone, but gathering data is particularly important for young scientists. It gives them a chance to see how previous data has been collected, and how studies work. Reading will not provide you this background or the insights associated with it, rather this knowledge only comes from observing the similarities and differences that arise across multiple subjects in an experiment.

So, be engaged as you run your study and then perform the analysis. These experiences can be a source for later ideas, even if you are doing what appears to be a mundane task. In addition, being vigilant can reduce the number and severity of problems that you and the lead investigator will encounter. Often, these problems may be due to changes in the instrument, or changes due to external events. For example, current events may change word frequencies for a study on reading. Currently, words such as bank, stocks, and mortgages are very common, whereas these words were less prevalent a few years ago. Billy Joel makes similar comments in his song "We didn't start the fire".

5.13 Running the low vision HCI study

While starting to setup to pilot, Judy identified the experiment's first major issue: the company's software was not cross-system compatible, it did not run on all versions of Windows. This was useful information, and helped refine the experimental setup and protocol.

During the pilot study, the two pilot subjects (who were legally blind and not part of the subject pool) identified persistent text-to-voice issues. The team was able to successfully implement a version of the software that was cross-system compatible for the experiment, but the text-to-voice issues could not be entirely eliminated within the time period allotted for the study.

These problems caused Judy to reconsider her test groups, adding two additional groups. Besides the control group (unmarked navigation bar) and the first experimental condition (marked navigation bar), she added two other experimental conditions: (a) a customizable graphical interface controlled through the arrow keys without a marked navigation bar, and (b) a customizable graphical interface with a marked navigation bar.

The decision to add a customizable graphical interface was in response to the text-to-voice issues—the company's text-to-voice processing had a difficult time with book and movie titles,

particularly if those titles included numbers. A major component of Judy's experiment tested the software's ability to support users browsing book and movie titles. The relative lack of surrounding text in these lists caused the software's hidden Markov models to frequently misread years as numerals. Because the software's statistical tools for disambiguating between differing pronunciations also largely depended on surrounding text, Judy's text-to-voice software would in some cases mispronounce words, failing to distinguish between the noun and verb forms of the word project for instance. Consequently in the pilot study, Judy was uncertain if the lag times associated with the original experimental conditions were, in fact, a result of the treatment or confusion caused by the text-to-voice issues.

To isolate to some extent the effects associated with the software, Judy's team implemented a customizable graphical interface that allowed users to increase the size of a selected object with the up and down arrow keys and the color with the left and right keys.

5.14 Running the multilingual fonts study

Developing our discussion from Chapter 4 regarding internal validity, we discuss specifically piloting. Through piloting, we often find procedural or methodological mistakes that have consequences for an experiment's internal and external validity. In the initial pilot data, Ying discovered a distribution in the data that she could not initially explain. The effect of changes in pixel density and size matched her expectations (denser letters were generally clearer as were larger ones, with the magnitude of these effects eventually flattening off. Also as expected, she did find a relationship between matrix formats and these thresholds when the participants encountered a black font on a white background. She, however, found that her color findings, even for Roman characters, did not match the literature. Previous work had shown that not only a font's size and density have an influence on its readability but also its brightness difference, and that light text on dark backgrounds and dark text on light backgrounds have predictably different distributions. Ying's and Edward's pilot data did not even remotely match the distributions found in the literature.

Ying and Edward began brainstorming about the possible causes for this discrepancy. Looking through the pilot study's screening questionnaire, Edward noted that there were no questions regarding color blindness. Further, the initial survey questions asked the participants to rank the matrix formats' colors relative to each other for formats of a given size and density. The initial list did avoid sequentially listing orange, red, and green matrix formats; however, it did list a blue matrix format followed by a yellow one. Many participants refused to complete the rankings because they could not see any distinguishable differences between the matrix format within a given size and density condition. Consequently, Ying's light background/dark font distribution was essentially bi-modal and incomplete, where the bi-modality was a result of whether the format was ranked or not.

To address this problem, Edward and Ying expanded the screening questionnaire to include questions about color blindness. In addition, they replaced their relative ranking scale and replaced it with a Likert scale, where participants encountered each color for a given condition separately. They then could respond to the question, "Do you find this sentence easy to read?" by selecting one of five answers: strongly agree, agree, unsure, somewhat disagree, or disagree. Summarizing the data required additional steps because we cannot assume the relative emotional distance between selections is constant—the distance between strongly agree and agree for instance may be larger or smaller than that between unsure and agree for a given topic. So, for the purposes of summarizing the data, Ying had to group selections into positive and negative responses and then order the color format within a given pixel/density condition with respect to the number of positive or negative responses collected. Ying could then see the gradation in user

preferences for the given brightness differences across the various matrix formats, both in new pilot data and later in the study.

5.15 Running the HRI study

A problem that Bob is very likely to find in running his study is that of recruiting suitable subjects. Unlike universities, companies frequently do not have a lot of people available. Often, the only people easily available are those that know about the product or who have a vested interest in seeing the product succeed commercially. These are not ideal subjects to test a robot. Bob will have to look into recruiting people through newspaper ads, casual contacts, and other contacts at and through the company.

In running his study in service of a company developing a product, Bob might find that he is most tempted to terminate his study or controlled observation early when he finds useful results. Of all our examples, it would be most appropriate for him to do this because that is what he is looking for, changes that lead to a better product. He is not looking for a general answer to publish, but is looking for results to improve his product. Now, if the people he is speaking to are hard to convince, he may particularly want to finish the study because robots are hard to set up and maintain, and more subjects pounding the table in frustration may be more convincing. Similarly, if he finds dangerous conditions or results that are conclusive on an engineering level, he has an obligation to provide his feedback early and to not put further subjects at risk.

5.16 Conclusion

Running the experiment is usually the culmination of a lot of work in developing the research question and hypotheses, planning the experiment, recruiting the subjects, and so on. It can also be the fun part, as you see your work coming to fruition and the data beginning to accumulate. There is lot to attend to while running an experiment, but it is the last step before you have data to analyze and find the answer to your research question.

5.17 Further readings

We can recommend a few resources for further reading.

- Huck, S. W., & Sandler, H. M. (1979). *Rival hypotheses: Alternative interpretations of data based conclusions*. New York, NY: Harper & Row.

  Rival hypotheses provides a set of one page mysteries about how data can be interpreted, and what alternative hypotheses might also explain the study's results. Following the mystery is an explanation about what other very plausible rival hypotheses should be considered when interpreting the experiment's results. This book is engaging and teaches critical thinking skills for analyzing experimental data. It also reminds you of biases that can arise as you run studies. It could be referenced in other chapters here as well.

- Mitchell, M. L., & Jolley, J. M. (2012). *Research design explained* (8 edition ed.). Belmont, CA: Wadsworth Publishing.

  Their appendix (Online practical tips for conducting an ethical study) has useful tips similar to this book.

5.18 Questions

Summary questions

1. Answer the following questions.
   (a) What is “debriefing”?
   (b) List the procedures for “debriefing” by Mitchell and Jolly (2007).
   (d) What is “a simulated subject”?

Thought questions

1. Refer to example scripts in Appendix B. Practice writing an experimental script based on your operational definitions of the research variables in Chapter 1.

2. Note how you would deal with the following potential problems in your study that you are preparing, or for one of the studies that has been used as an example: a subject becoming ill in the study, a subject becoming lost and arriving 20 minutes late with another subject scheduled to start in 10 min., a subject coming in an altered state, a subject self-disclosing that they have committed an illegal act on the way to the study, a subject that discloses orally that private medical history, a subject that discloses on a study form private medical history.

6 Concluding a Research Session and a Study

This chapter provides practical information about what you should do when you get done with your experiment.

6.1 Concluding an experimental session

6.1.1 Concluding interactions with the subject

After your subject has finished participating in your experiment, there are important parts of your interaction with him or her left to complete. The first of these is *debriefing*. As discussed in Chapters 3 and 5, if your study has involved deception, you must usually reveal this deception to the subject. Even if there was no deception, it is good practice to spend a few minutes debriefing the subject about the purpose of the study—your hypotheses, how you hope to use the results, and so on. These are things you generally don't want to mention at the beginning of the experimental session, but that will help your subject understand the value of their participation. The second topic is providing compensation, whether that is course credit or monetary payment. Handling this appropriately is important both for leaving subjects with a good impression of your study and for maintaining records you may need for your research sponsor or department.

It is also wise when concluding the experiment to make sure that you have all of the information you need from the subject. Do you have your copy of the consent document signed by the subject? Is information that will allow you to link pencil-and-paper data with computer data files properly recorded?

6.1.2 Verifying records

After each subject, it is a good idea to check to make sure that data files are properly closed. For example, if an EPrime program is terminated not by running to its normal conclusion but by shutting down the computer, the data file may not be saved correctly. Any paperwork, whether it contains data (for example, a questionnaire) or simply clerical work (how much credit should be given) should be verified and appropriately filed.

This is also an appropriate time to anonymize the data, as discussed in Chapter 3. You will of course want to retain a record of subjects' names for purposes of assigning credit or documenting payment, but if it is not necessary to associate their name with their data, it should be removed. Depending on the nature of the research, you may want to store a list of subject codes and names that could later be used to re-link identity information with the data, but you should consider carefully whether this is necessary. It is also useful to keep notes about every subject. For example, if something unusual happened—the subject reported an apparent problem with the experimental software, the subject seemed to ignore instructions, a loud distraction occurred in the hallway—this should be noted, so that the lead researcher or principal investigator can make a judgment about whether to include that subject's data, to conduct additional tests on the software, etc. Don't think "I'll remember to mention this at the lab meeting"—trust us, you won't, at least some of the time. One of us asks our research assistants to initial a list of subjects to verify that everything went smoothly, including entering the correct information in the program running the experiment, starting on time, and so on. Sometimes, too, a subject will say something that provides an insight into the research question—if that happens, write it down at the end of the session. Such insights can be like dreams: clear and vivid in the moment, and impossible to remember later.

It is also useful to document, perhaps in a lab notebook, information such as the date that particular data were collected (the dates on data files may reflect when they were last accessed rather than when they were collected), the file names for programs used to collect data, and so on.

This advice may seem obsessive, but it comes from long experience in running experiments. It is likely that the experiment you are running is one of many conducted in the laboratory you're working in, and perhaps one of many that you're running yourself. Having a record you don't need is not a problem; lacking a record you do need may mean that the data collection effort was wasted, or at least that you will need to spend a lot of time reconstructing exactly what you did.

6.2 Data care, security, and privacy

All information and data gathered from an experiment should be considered confidential. If others who are not associated with the experiment have access to either data or personal information, the participants' privacy could be violated. Thus, it is the responsibility of lead researchers and experimenters to ensure that all security assurance procedures are explained and enforced.

Researchers must safeguard against the inappropriate sharing of sensitive information. Personal information about the participants must not be shared with people not associated with the study. Thus, the data should not be left untended. In most studies, experimental data are kept in locked files or on secure computers. The level of security may vary with the type of data. Anonymizing the data (removing personally identifying information) is a strong protection against problems. Anonymous reaction time data, where the only identifying information is a subject ID, is low or no risk. Personal health records where the subjects might be identified are much more sensitive, and would require more cautious storage, perhaps being used only on a removable disk that is locked up when not in use.

6.3 Data backup

To protect against data loss, back up all of your data routinely (after running a subject, and every day when you are doing analyses of the data). If your data is stored in electronic files, store them in a secure hard drive or burn them onto a CD. If you are using paper documents, they can be scanned and stored on a computer file as back up. We suggest that you back up your data after each subject rather than weekly while conducting a study.

6.4 Data analysis

If you have planned your data collection carefully, and pilot-tested your data collection and analysis plans, the data analysis stage of your experiment may be straightforward. However, there are often complexities, especially if you are dealing with a complex data set with multiple independent and dependent variables, or complex measures such as verbal protocols. Even with carefully planned analyses, additional questions often arise that require further data analyses. If your research is submitted for publication, editors and reviewers may ask for additional analyses.

6.4.1 Documenting the analysis process

Our advice is to document the data analysis process very carefully. Many, perhaps most, experiments will require that you transform the data to analyze it. For example, if you have within-subjects variables, you will usually need to aggregate and transform the data so that levels of your independent variable are represented as columns rather than the rows likely to be in your data file. These transformations may be done in the program you use for analysis (e.g., SPSS) or in a spreadsheet program such as Excel. Keep careful notes on the steps of transformation, and

never—never!—discard the original, untransformed data files. If you filter your data to remove subjects who didn't follow instructions, outlying data points, etc., keep a record of exactly what you did, and the names of filtered and unfiltered data files.

We find it is often useful to summarize your results as you work through the data analyses. The goal of data analysis is not to mechanically work through statistical procedures, but rather to understand your data set and what it can tell you. It is useful to look not just at means—though differences in means may be most important for your hypotheses—but at the actual distributions of data, how much they vary, and so on. Experienced researchers learn how to evaluate their data and analyses for plausibility—if something seems "off," it might be due to anomalous data (perhaps caused by a subject not taking the task seriously, an error in data entry, etc.), an error in manipulating the data file, or some other study related reason. Thinking about whether the results of your analysis make sense, and understanding how problems with the data can be responsible for odd results in your analysis is important.

It is likely that your data analysis will result in a number of output files. While most statistical software provides output that will let you trace exactly what you did to generate the output, doing so can be time-consuming. Keeping good notes of what is in each output file is likely to save time in the long run.

6.4.2 Descriptive and inferential statistics

Your data analysis will include two kinds of statistics, descriptive and inferential. Descriptive statistics are those that, as the name suggests, *describe* your data. Means and other measures that show the average or typical value of your data, standard deviations and other measures that show the variability of your data, and correlations and regressions that show the relations among variables are all descriptive statistics. Statistics texts will define a variety of descriptive measures, and statistical software will calculate many measures. When you are working with your own data you will come to understand how important the choice of descriptive statistics can be. Does the mean really reflect the typical value of your data? Is the standard deviation misleading because your data includes many extreme data points? The decisions you make about descriptive statistics are choices about the best way to summarize your data, both for your own thinking and to communicate to others.

Detailed advice on choosing descriptive statistics is beyond the scope of this book. However, we can offer this general advice—explore the possible descriptive statistics so that you get to know your data set. We have often seen researchers who missed or misconceived aspects of their data because they failed to consider a variety of ways to summarize their data. Considering multiple graphic depictions of your data can be very useful—for example, looking at a distribution of response times may immediately show that the mean is not a good representation of the typical response time, perhaps suggesting that the median, which is less sensitive to extreme values, would be a better description. Graphing means, especially in factorial experiments that might reveal interactions among variables, can visually communicate trends that are difficult to see in a table of means. One of us has several times had the experience of re-graphing an interaction that a research assistant thought was uninterruptible, and immediately finding a simple and meaningful description of the interaction.

A particular issue that arises in describing data results from aggregating data over subjects. Of course we want to know what is common in the performance of all of our subjects (within an experimental condition) taken together. Sometimes, though, averaging over subjects results in a very misleading picture of what actually happened. A recent example with important theoretical and practical implications concerns learning curves observed in studies of skill acquisition. For many skills, performance as measured by response time improves with practice following a

power function, such that performance speeds up very quickly over the first few trials of practice, then continues to speed up more slowly with additional trials (Crossman, 1959; e.g., Seibel, 1963). Newell and his colleagues were so impressed by this finding that they proposed the *power law of practice* as a basic phenomenon of skill acquisition (Rosenbloom & Newell, 1987). However, other researchers have pointed out that a power function for speedup can result from averaging over subjects, none of whom individually show such a speedup. For example, if each subject discovers a strategy that results in a dramatic, sudden, single-trial speedup, but individual subjects discover the strategy after different numbers of trials, the average may suggest a gradual, power-function speedup displayed by no individual subject (Brown & Heathcote, 2003; Delaney, Reder, Staszewski, & Ritter, 1998; Estes, 1956).

Table 6.1 and Figure 6.1 illustrate this point using hypothetical data. Imagine subjects learning to perform a simple task that takes 1,000 ms (one second) to complete, until you find a shortcut that allows you to complete the task in half the time (500 ms). If subjects vary in when they discover the shortcut, as illustrated in Table 6.1, averaging response time data over subjects will generate the data points displayed in Figure 6.1. The dashed line shows the best-fitting power function for these data points. Examining the graph suggests that learning is a smooth, continuous process with a power-law speedup over trials. However, the actual process is a sudden discovery of a shortcut, resulting in a sharp, step-function speedup in performance. Thus the averaged data obscures the true form of learning.

Table 6.1. Response time in milliseconds by learning trial (hypothetical data). Italics indicate the first trial after discovering a shortcut, bold indicates the last trial before discovering the shortcut.

Subject	**Learning trial**							
	1	2	3	4	5	6	7	8
1	**1,000**	*500*	500	500	500	500	500	500
2	**1,000**	*500*	500	500	500	500	500	500
3	**1,000**	*500*	500	500	500	500	500	500
4	**1,000**	*500*	500	500	500	500	500	500
5	**1,000**	*500*	500	500	500	500	500	500
6	**1,000**	**1,000**	*500*	500	500	500	500	500
7	**1,000**	**1,000**	*500*	500	500	500	500	500
8	**1,000**	**1,000**	*500*	500	500	500	500	500
9	**1,000**	**1,000**	**1,000**	*500*	500	500	500	500
10	**1,000**	**1,000**	**1,000**	**1,000**	*500*	500	500	500
11	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	*500*	500	500
12	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	*500*	500
Mean	**1,000**	**791**	**667**	**625**	**583**	**542**	**500**	**500**

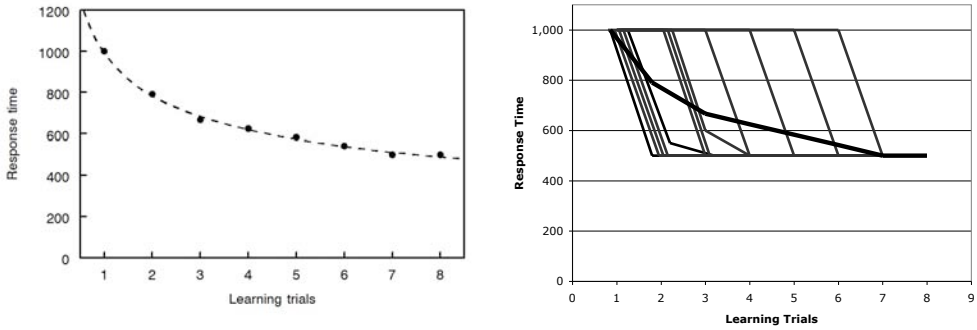


Figure 6.1 Mean response time as a function of trial, with power law fit (data from Table 6.1) (left), and the individual learning curves (right) superimposed on the average response time.

Of course you want not just to describe your data, but to draw conclusions from it. This is where *inferential* statistics come into play. Again, detailed discussion of the many possible inferential statistics is beyond the scope of this book. However, we can offer a few pieces of advice. The first is to make sure that the statistics you choose are appropriate for your data and your research question. Most researchers have learned to use analysis of variance as their primary analysis tool. However, when independent variables are on interval or ratio scales (see Chapter 2), regression analyses and their associated inferential statistics may be much more powerful. For example, it has become common in several areas of psychology to use working memory capacity as an independent variable, dividing subjects into those with high and low (above or below the median) capacity. ANOVA *can* be applied to such data, but does not provide the most powerful test of whether working memory capacity affects the dependent variable. A second piece of advice is not to fall in love with a particular description of your data until you know that inferential statistics support your interpretation. Contrary to what some researchers think, inferential statistics do not draw your conclusions for you—instead, they tell you which of your conclusions are actually supported by the data. Third, and finally, don't fall into the trap of null-hypothesis reasoning—believing that the failure to find a significant difference is equivalent to finding evidence that there is no difference. Many research articles have been rejected for publication in part because a researcher argued that there was a meaningful difference in performance in one case, and equivalent performance in another, when the statistical tests simply fell on either side of the conventional criterion for statistical significance.

6.4.3 Planned versus exploratory data analysis

If you have followed the advice in this book, you planned your data analyses well in advanced. If that planning was successful, following the plan should provide the answers to your research questions and evidence for or against your hypotheses. However, most data sets are complex enough to allow additional analyses that are exploratory rather than planned. For example, your main question may be which of two experimental conditions resulted in greater accuracy. However, your data may allow you to explore the question whether subjects' performance was more variable in one condition than another, or whether their performance depended on the order in which they solved problems (even though counterbalancing meant that order did not affect the test of your main hypothesis). Exploratory data analyses are often the source of additional insights into the research question, or the basis of ideas for additional experiments.

6.4.4 Displaying your data

If you are analyzing data, you will eventually need to communicate your results to someone—perhaps the principal investigator supervising your research, perhaps colleagues at a conference, perhaps the editors, reviewers (and, one hopes, readers) of a journal. The diversity of possible research results makes it difficult to give general advice; but during the data analysis stage, we have one important suggestion: make pictures, early and often. A graph can make apparent features of your data that are hard to extract from a data analysis output file. If you have data that can be graphed in more than one way, do so. Modern software tools make it easy to generate graphs, and graphs are usually a much more efficient way to communicate your results—even to yourself—than tables or lists of means.

6.5 Communicating your results

Rarely does anyone run an experiment only for their own information. Instead, one of the goals is usually to communicate the results to others. In this section, we discuss some considerations about sharing your results.

6.5.1 Research outlets

The written product resulting from your research project may take several forms. One, a *technical report*, is usually written primarily as a report to the research sponsor. A technical report may be a final summary of a project, or serve as a progress report on an ongoing project. Technical reports are often written in a format specified by the research sponsor.

Another possibility is a presentation at a scientific conference. Such presentations may take several forms. One form is a talk, usually accompanied by slides prepared with PowerPoint or similar software. A typical conference talk is 10 to 20 minutes in length, followed by a brief question-and-answer session. Another kind of conference presentation is a poster. At a poster session, there are dozens, or sometimes hundreds, of posters in a large hall of some kind, with poster authors standing at their posters to discuss their research with individuals who stop by. Conference presentations can be very useful for getting feedback on your research, which can be helpful in preparing an article for publication or in planning future experiments. Sometimes conference presentations are accompanied by brief papers published in the proceedings of the conference.

Often, the final product of a research project is an article in a scientific journal. For researchers employed in—or aspiring to—academic settings, a journal article is usually the most valuable product for advancing a career. A major benefit of publishing research in a scientific journal is that it will be available for other researchers. Scientific journals are typically quite selective in choosing articles to publish; some reject as many as 90% of the articles submitted for publication. Frequently, however, an initial rejection is accompanied by an invitation to revise and resubmit the article for reconsideration, or by suggestions for additional data collection to increase the value of the research to the field. Students sometimes ask us if researchers are paid for articles published in scientific journals—the answer is no, they are not. Scientific journals are a means by which researchers in a field communicate to one another, including to future researchers.

Reporting research results at a conference or in a scientific journal involves some form of *peer review*. This means that an editor or a conference program committee receives comments from several researchers with expertise in the area of research, and uses those comments to decide whether to accept the proposed presentation or article. For a conference, the reviewers may consider a brief summary of the presentation, or a more complete paper to be published in conference proceedings. The peer review process can seem quite daunting, but if you take the

comments of reviewers as feedback on how to improve your research or your presentation of it, you will find it to be quite helpful.

Choosing an outlet for your research will depend on several factors. Technical reports are usually mandated by the grant or contract that serves as an agreement with the research sponsor. A conference presentation may be an appropriate outlet for exploratory or partially completed research projects that are not yet ready for scientific journals—"works in progress" are appropriate for many conferences but usually not for scientific journals. Decisions about outlets are usually made by the lead researcher or principal investigator.

Regardless of the outlet used to communicate your research, it is important to adjust your writing to the specific goals of and audience for the outlet. A technical report, for example, may have as its audience employees of the research sponsor who are not themselves researchers, and a primary goal may be recommendations concerning the practical implications of the research. The audience for a conference presentation is usually researchers working on similar topics, and the presentation should be developed with that audience in mind. A short conference talk or a poster make efficiency of communication very important, and your goals may include having people remember your main message or making clear the aspects of the research for which feedback would be helpful. The audience for a journal article is also researchers in related fields, but it useful to keep in mind both the general style of the journals you have in mind and your own experience as a reader of journal articles.

6.5.2 The writing process

Guidance on writing research reports is beyond the scope of this book, but the Publication Manual of the American Psychological Association describes the standard format for preparing manuscripts for publication in psychological journals, and offers some guidance on writing. We can say that if you have followed the advice in this book about preparing to run your study and keeping records, you will find yourself well prepared to begin writing.

One piece of advice we have concerning writing is this: Do not expect your first—or second, or third—draft of your research report to be final. Members of a research team will pass drafts of research reports back and forth, and it is not unusual for there to be five, six, or even more drafts of a paper before it is submitted to a journal. And once an article is submitted to a journal, it is almost certain that it will have to revised in response to comments from the editor and reviewers. We have found that this aspect of writing about research is often difficult for students to accept and appreciate—they are used to submitting class papers after one or two drafts. One consequence of this is that student researchers are often reluctant to share their early drafts with others, including advisers or principal investigators. This is a mistake—our best advice to student researchers is to share your work early and often, with anyone who is willing to read it and provide comments. Your goal should be to effectively communicate your research, not to impress others with your ability to produce a polished, finished product on the first try.

6.6 Concluding the low vision HCI study

As Judy did the analyses and wrote up a short report summarizing the study, she found that the marked navigation bar with the customizable interface had the lowest lag times for the majority of users, followed by the customizable interface with an unmarked navigation bar, and the marked navigation bar with no customizable interface. As expected, the control condition had the longest lag times. The study's sample size made it impossible to determine if the marked navigation bar had a significant effect for participants unable to detect to changes in color or size (n=1). For participants able to distinguish to some extent size or color (n=31), the difference between the control group and fourth condition (the combination of a customizable interface and

a marked navigation bar) was statistically significant, indicating that marked navigation bars do have a complimentary effect. The differences between the other three conditions followed the expected trend but were not statistically significant.

There remains the question of the software. With better software, the relative differences between the four conditions might differ. Regardless, we would expect the lag times to decrease. As for the marked navigation bar's impact on performance for participants with virtually no visual acuity, Judy will need to find more participants. In addition, future studies are necessary to see if these trends translate to portable devices. E-readers and tablets are only now beginning to routinely support text-to-voice processing. Yet, they and tablets are important new markets for her company.

To conclude, Judy's experiment is not unusual. Her study uncovered important trends but requires further studies to fully understand and extend these findings. Nevertheless, Judy findings can be usefully incorporated in future products.

6.7 Concluding the multilingual fonts study

Edward and Ying's experiences provide some insights for both concluding study sessions and concluding studies. Also, they provide an example of how to skillfully transition a technical report required for grant stakeholders (the One Laptop Per Child Project, in this case) into a conference paper. The diverse cultural backgrounds and experiences of the study's participants made debriefing more important than is sometimes the case in HCI studies. In many cases, participants from outside the immediate university community volunteered to participate in the study because of the study's connection to the OLPC. While compensated, they, nevertheless, did this often at some personal discomfort, coming to a new place and interacting with people in a non-native language. Frequently, their reading comprehension far exceeded their verbal fluency, making the experiment easy to complete but getting to the experiment and understanding the initial instructions more difficult. In this case, debriefing provided a way to not only go over what happened in the experiment but also to thank the participants and show how their participation was contributing to an important goal, literacy.

Like most studies, Ying's and Edward's study highlighted new research questions, as well as contributing new findings. Ying and Edward did find reliable differences in the preferences of users across the experimental conditions. As expected, they also found that users generally preferred darker fonts on lighter backgrounds. On the other hand, there are further questions. For instance, while this study suggested that 8x8 formats were preferable, the pedagogical literature suggests that children do respond more favorably to greater brightness differences than most adults. This work, however, has generally occurred in the United States. Other studies (Al-Harkan & Ramadan, 2005) suggest that color preferences at least seem to be culturally relative. Therefore, testing the generalizability of the findings from the pedagogical literature and how they might inform user interface design for OLPC users requires more studies. Ying had considered this problem early in the experimental design process after uncovering these findings during the literature review; but when recruiting proved to be difficult, Ying and her adviser determined that recruiting enough child participants would be infeasible. Noting in the discussion sections of both the technical report and conference paper the need for this follow-up study, Ying proposed an on site or web-based study at various OLPC locations, especially since the screen resolution would already be consistent and this research question seems to entail fewer environmental controls.

Moving this work from a technical report to a conference paper was a relatively simple process with respect to writing new content, with the most difficulty associated with focusing the presentation to a few key ideas. Initially, Edward had a difficult time identifying and

summarizing key procedural details. In addition, allowing Ying and the PI to see his incomplete fragmentary work was a difficult process. Fortunately, by this time, Edward trusted Ying enough to submit these ugly drafts and take advice from her on his writing. Looking back, having the tech report in hand made this process far easier.

Nevertheless, learning to work through multiple drafts (including updating the date and version number for each draft), managing references, and finding a weekly meeting schedule that met everyone's needs required some patience and negotiation. In our experience, we find these are common problems for undergraduate and graduate students working on their first few publications. Have patience and take the advice of your co-authors with an untroubled heart—there will be a lot of revisions but they are a normal part of the process.

6.8 Concluding the HRI study

There are aspects of Bob's work that influence how to wrap up a study. The first is that someone will care what he finds. Bob should find out the norms and style of result summaries that will influence the engineers and managers the most, and provide the results to them in these formats. This may include a choice of what is reported (speed, accuracy, satisfaction) and be in the format of a short email, this may be a short tech report, and it may in some cases be edited videos of subject's experiences using the robot. He should keep in mind that his goal is to improve the product and to present his results in ways that are easy for the engineers to understand and to act upon.

The second assumption is that he may not be the person to make use of the data at a later time. As Bob wraps up his study he should be careful to anonymise his results so that the privacy of his subjects will remain protected. He should label his results, either with paper in a folder or in a computer file associated with the data and analyses files. He should with the time allowed to him document what he did and what he found. He should archive this as best he can at the company, perhaps in a company library if there is one, or with his manager and the technology developer. In some ways his results will come out as a report like the other example projects, because that style of report is useful and such reports will make the data and results more understandable across time and distance.

Bob particularly needs to keep in mind how to be persuasive. The British Psychological Society (*The Psychologist*, 24(3), p. 179) summarized it very well: "the best way chance of changing the minds of non-believers would be an artful combination of clear, strong logical argumentation mixed with value-affirming frames and presented in a humble manner that produces positive emotional reaction." So, Bob must make his argument for changes clearly, using the values and ethics shared by the company (for example, acknowledging the technical achievement and also noting the costs for change); this too means writing well, broadly defined to include well done and appropriate figures and formatting in a style the readers expect.

6.9 Conclusion

As we discussed in Chapter 1, finishing an experiment is often not the end of the research process. Often, the results will lead to new or refined research questions, better experimental tasks, designs, or procedures, or all of these. Nevertheless, there is great satisfaction in completing an experiment and seeing the results. This is the payoff, and it is important to make sure that you wrap things up effectively.

6.10 Further readings

There are few materials on how to finish a session of a study, but there are plenty of materials on how to anlayse your data and communicating your results.

- Howell, D. C. (2008). Fundamental statistics for the behavioral sciences (6th ed.). Belmont, CA: Thompson Wadsworth.

  This is one of several good, commomly used statistics books.

- Huff, D., & Geis, I. (1993). *How to lie with statistics*. New York, NY: W.W. Norton.

  This, and ealier versions, discuss how to interpret results and report them.

- Strunk, W., & White, E. B. (1979 or any edition). *The elements of style*. NY, NY: Macmillan.

  This is a timeless first book on how to improve writing.

6.11 Questions

Summary questions

1. How does running the sessions provide you with a "chance for insights"? Can you think of or find an example of this happening?

2. In completing your research study, the final product can be such things as an article in a journal or in a conference. Reporting results may involve a form of "peer review". Describe what "peer review" is.

Thought questions

1. It cannot be over emphasized that data backup is important in conducting a research study with subjects. Discuss how you plan on data backup in your research.

2. Given an ideal world, where would you suggest that the researchers in the three examples (Low vision HCI study, Multilingual fonts study, and HRI study) publish their work?

7 Afterword

There are many books available about research methods and related statistical analyses. We, however, realized that students usually do not have a chance to learn how to run their own experiments, and that there are no books that we are aware of that teach students practical knowledge about running experiments with human participants.

Students charged with running experiments frequently lack specific domain knowledge in this area. Consequently, young researchers chronically make preventable mistakes. With this book, we hope to assist students as they begin to obtain hands-on knowledge about running experiments. The topics and guidance contained in this book arise from the authors' collective experience in both running experiments and mentoring students.

Further methods of gathering data are being developed. Though these changes will impact the development of future experimental procedures, the gross structures of a study and the aspects we have discussed here, of piloting, scripts, anonymizing data, and so on, are not likely to change.

As you venture into research, you will find new topics that will interest you. In this text, we are not able to examine all populations or touch upon measurements and tools that require additional training. Consequently, we are not able to cover in detail the collection of biological specimens, eye-tracking, or fMRI. However with further reading and consultation with colleagues, you will be able to master these skills.

Running studies is often exciting work, and it helps us understand how people think and behave. It offers a chance to improve our understanding in this area. We wish you good luck, bonne chance, in finding new scientific results.

Appendix 1: Frequently Asked Questions

How do I know my study measures what I want it to?
See:

How do I start to plan a study?
See:

Do I need to get IRB for my work?
See:

What should I do if I don't need to get IRB?
See:

Glossery of terms as well?

Independent variable	A variable that is manipulated in the study, either by assignment of materials or assignment of subjects.
Dependent variable	A measurement that is taken during the study, such as reaction time, or percent correct. It depends on other things.
Pilot study	An abbreviated version of the study done to test the procedure and prepare for a larger study.
Power	The power in an experimental study indicates the probability that the test (or experiment) will reject a false null hypothesis. Failure to reject the null hypothesis when the alternative hypothesis is true is referred to as a Type II error. Thus, as the power of a study increases, the chances of a Type II error decrease.
IRB	Internal Review Board. They review study proposals to ensure safety and compliance with US federal regulations.
Informed consent form	
Null hypothesis	The hypothesis that the treatment DOES NOT lead to differences. For example, the null hypothesis might be that two interfaces are equally easy to use.

Appendix 2: A Checklist for Setting up Experiments

This checklist contains a list of high level steps that are nearly always necessary for conducting experiments with human participants. As an experimenter or a principal investigator for your project, you need to complete the items below to set up experiments. You might use this list verbatim or you might modify it to suit your experiment. The list is offered in serial order, but work might go on in parallel or in a different order.

- ☐ Identify research problems and priorities, design experiment
- ☐ Prepare the IRB form and submit it to the office of research protection, noting how to address any harm or risks
- ☐ Prepare "consent form"
- ☐ Prepare "debriefing form"
- ☐ Set up the experiment environment
- ☐ Run pilot tests to check your experimental design and apparatus
- ☐ Analyze pilot study data
- ☐ Prepare experiment script
- ☐ Receive IRB approval
- ☐ Advertise the experiment and recruit participants (e.g., a flyer, a student newspaper)
- ☐ Run the experiment
  (Make sure a lab for the experiment is available for when you need to run)
  - Explain the experiment to participants (e.g., purpose, risk, benefits)
  - Gather data and store data
- ☐ Report results

Appendix 3: Example Scripts to Run an Experiment

High level script for an HCI study

This is a short example script. While experiments will differ, this script includes many common elements. It was used for Kim's PhD thesis study (J. W. Kim, 2008).

Experimenter's Guide

This is an example high level summary script for an experiment. Every experimenter should follow the procedures to run a user study about skill retention.

(1) Check your dress code

(2) Before your participants are coming in, you need to set up a set of the experiment apparatus.

Start RUI in the Terminal Window. (see details ..)

Start the Emacs text editor.

Prepare disposable materials, handouts, such as informed consent form

(3) Welcome your participants

(4) Put a sign on the door indicating that you are running subjects when the experiment starts

(5) Give the IRB approved consent form to the participant and have them read it

(6) If they consent, start the experiment

(7) Briefly explain what they are going to do

(8) Give them the study booklet.

Participants can use 30 min. maximum to study the booklet.

(9) While participants are reading the booklet, you can answer their questions about the task.

(10) Turn on the monitor that is located in the experimental room, so that you can monitor the participant outside the room.

(11) When the experiment is finished, give an explanation about the payments or extra credit. Thank them; give them a debriefing form. Also, if there are any additional schedules for later measures, remind them.

(12) Take down the sign on the door when the experiment is done

(13) Copy the data to the external hard drive

(14) Shut down apparatus

(15) Make supplies for the next subject

Using RUI

RUI (Recording User Input) will be used to log keystrokes and mouse actions of the participant. RUI requires Mac OS X 10.3 (Panther) or later versions. It has been tested up to Mac OS X 10.5.8 (Snow Leopard). For RUI to record user inputs, "Enable access for assistive devices" must be enabled in the Universal Access preference pane.

(1) Launch Terminal

(2) In Terminal, type the below information:
./rui –s "Subject Name" –r ~/Desktop/ruioutput.txt
(3) You will get this message:
rui: standing by—press ctrl+r to start recording…
(4) Press "CTRL+r"
(5) To stop recording, press "CTRL+s"

Note:
If you see the message of "-bash: ./rui: Permission denied" in the Terminal window, you need to type "chmod a+x rui" while you are in the RUI directory.

Measuring Learning and Forgetting

Emacs is started by the experimenter for every session. The participants will start and stop RUI to record their performance. The experimenter needs to ensure that the participants cannot do mental rehearsal during the retention period.

More detailed script

This script was used in conducting an experiment reported in Carlson and Cassenti (2004).

1. Access the names of participants from subject pool. Go to subject pool under "favorites" in Explorer, type in experiment number 1013 and password ptx497. Click on the button labeled "view (and update) appointments." Write down the name of participants on the log sheet before they start arriving.
2. Turn on computers in subject running rooms if they aren't already on. If a dialog box comes up asking for you to log in, just hit cancel
3. As participants arrive, check off their names on your list of participants. Make sure that they are scheduled for our experiment – sometimes students go to the wrong room.
4. Give each participant two copies of the informed consent (found in the wooden box under the bulletin board). Make sure they sign both copies and you sign both copies. Make sure to ask if the participant has ANY questions about the informed consent.
5. Fill out the subject running sheet with subject's FIRST name only, handedness (right or left), gender, the room in which he or she will be run, and your name.
6. Begin the experiment by click on "simple counting" file on desktop. Once the program opens press F7. Enter the subject number from the subject running sheet when it asks for session number you should always enter "1." Double check the information when the confirmation box comes up. If the next screen asks you if it's okay to overwrite data, click "no" and put in a different subject number, changing the experiment sheet as needed. If you want to do all of this while the participant is reading the informed consent to save time go right ahead, but make sure to answer any informed-consent related questions the participant may have.
7. Take the participant to the room and say the following: "**This experiment is entirely computerized, including the instructions. I'll read over the instructions for the first part of the experiment with you."** Read the instructions on the screen verbatim. Ask if they have any questions. After answering any questions they may have, leave the room and shut the door behind you. Place the "Experiment in Progress" sign on the door.
8. At two points during the experiment subjects will see a screen asking them to return to room 604 for further instructions. When they come out you can lead them back to room taking along the paper for Break #1 and a pen. Read aloud to them the instructions that

are printed on the top of the sheet and ask if they have any questions. Give the participant two minutes to work on their list then come back in and press the letter "g" (for **g**o on). This will resume the experiment where they left off. Ask again if they have any questions, then leave the room again and allow them to resume the experiment. The second time the subject returns to 604 follow the same procedure this time with the instructions and paper for Break #2.

9. Fill out the log sheet if you haven't done so. You should have the necessary information from the subject pool. If somebody is signed up but doesn't show up, fill out the log sheet for that person anyway, writing "NS" next to the updated column.
10. Fill out a credit slip for each participant, and be sure to sign it.
11. Update participants on the web. Anyone who doesn't show up (and hasn't contacted us beforehand) gets a no show. People who do show up on time should be given credit. If they come too late to be run you may cancel their slot.
12. Participants should leave with three things: a filled out credit receipt, a signed informed consent from, and a debriefing. Ask them if they have any other questions and do your best to answer them. If you don't know the answer you can refer to Rachel or Rich (info at the bottom of debriefing). Make sure to thank them for their participation
13. When done for the day, lock up subject running rooms (unless someone is running subjects **immediately** after you are and is already there when you leave). If you are the last subject runner of the day please turn off the computers. Always lock up the lab when you leave unless someone else is actually in the lab.

Appendix 4: Example Consent Form

Here is an example of an informed consent form that you can refer to when you need to generate one for your experiment. This is taken from Kim's thesis study (J. W. Kim, 2008).

ORP USE ONLY: IRB#21640 Doc. #1
The Pennsylvania State University
Office for Research Protections
Approval Date: 09/09/2008 – J. Mathieu
Expiration Date: 09/04/2009 – J. Mathieu
Biomedical Institutional Review Board

Informed Consent Form for Biomedical Research
The Pennsylvania State University

Title: Investigating a Forgetting Phenomenon of Knowledge and Skills

Principal Investigator: Dr. Frank E. Ritter
316G IST Bldg, University Park, PA 16802
(814) 865-4453 frank.ritter@psu.edu

Other Investigators:

Dr. Jong Wook Kim
316E IST Building
University Park, PA 16802
(814) 865-xxx; jongkim@psu.edu

Dr. Richard J. Koubek
310 Leonhard Building
University Park, PA 16802
(814) 865-xxxx rkoubek@psu.edu

1. **Purpose & Description:** The purpose of the study is to investigate how much knowledge and skills are forgotten and retained in human memory after a series of learning sessions. Human performance caused by forgetting will be quantitatively measured. If you decide to take part in this experiment, please follow the experimenter's instruction.

   The experiment is held at 319 (Applied Cognitive Science Lab.) or 205 (a computer lab) IST building. During the experiment, the timing of keystrokes and mouse movements will be recorded.

   A group of participants (80 participants) selected by chance will wear an eye-tracker to measure eye movements during the task, if you consent to wear the device. You can always refuse to use it. The eye-tracker is a device to measure eye positions and eye movements. The eye-tracker is attached to a hat, so you just can wear the hat for the experiment. The device is examined for its safety. You may be asked to talk aloud while doing the task.

2. **Procedures to be followed:**
   You will be asked to study an instruction booklet to learn a spreadsheet task (e.g., data normalization). Each study session will be 30 minutes maximum. For four days in a row, you will learn how to do the spreadsheet task.

   Then, you will be asked to perform the given spreadsheet tasks on a computer (duration: approximately 15 minutes).

With a retention interval of 6-, 9-, 12-, 18-, 30-, or 60-day, after completing the second step, you will be asked to return to do the same spreadsheet task (duration: approximately 15 min/trial)

3. **Voluntary Participation:** The participation of this study is purely based on volunteerism. You can refuse to answer any questions. At any time, you can stop and decline the experiment. There is no penalty or loss of benefits if you refuse to participate or stop at any time.

4. **Right to Ask Questions:** You can ask questions about this research. Please contact Jong Kim at jongkim@psu.edu or 814-865-xxx with questions, complaints, concerns, or if you feel you have been harmed by this research. In addition, if you have questions about your rights as a research participant, contact the Pennsylvania State University's Office for Research Protections at (814) 865-1775.

5. **Discomforts & Risks:** There is no risk to your physical or mental health. You may experience eye fatigue because you are interacting with a computer monitor. During the experiment, you can take a break at any time.

6. **Benefits:** From your participation, it is expected to obtain data representing how much knowledge and skills can be retained in the memory over time. This research can make a contribution to design a novel training program.

7. **Compensation:** Participants will receive monetary compensation of $25, $30, or $35 in terms of your total trials, or extra credits (students registered to IST 331). The experiment consists of 5 to 7 trials ($5 per trial). The compensation will be given as one lump sum after all trials. For the amount of $30 and $35, participants will receive a check issued by Penn State. Others will receive a cash of $25. Total research payments within one calendar year that exceed $600 will require the University to annually report these payments to the IRS. This may require you to claim the compensation that you receive for participation in this study as taxable income.

8. **Confidentiality:** Your participation and data are entirely confidential. Personal identification numbers (e.g., PSU ID) will be destroyed after gathering and sorting the experimental data. Without personal identification, the gathered data will be analyzed and used for dissertation and journal publications. The following may review and copy records related to this research: The Office of Human Research Protections in the U.S. Department of Health and Human Services, the Social Science Institutional Review Board and the PSU Office for Research Protections.

You must be 18 years of age or older to take part in this research study. If you agree to take part in this research study and the information outlined above, please sign your name and indicate the date below.

You will be given a copy of this signed and dated consent for your records.

_____ _____

Participant Signature Date

_____ _____

Person Obtaining Consent (Principal Investigator) Date

Appendix 5: Example Debriefing Form

[This is the debriefing form used in the study reported in Ritter, Kukreja, and St. Amant (2007).]

Human-Robot Interaction Study Debriefing Form

Thank you for participating in our human-robot interface testing study.

From your participation we will learn how people use interfaces in general and Human-Robot interfaces in particular. These interfaces are similar to those used to interfaces used to work in hazardous areas including those used in rescue work at the World Trade Center. By participating, you have been able to see and use a new technology. The results can lead to improved interfaces for robots that replace humans in hazardous conditions.

You may also find the Robot project overview page useful and interesting.

If you have any questions, please feel free to ask the experimenter. You can also direct questions to Dr. Frank Ritter, (frank.ritter@psu.edu, 865-4453).

Appendix 6: Example IRB Application

Your Internal Review Board will have its own review forms. These forms are based on each IRB's institutional history, and the types of studies and typical problems (and atypical problems) that they have had to consider over time. Thus, the form we include here can only be seen as an example form. We include it to provide you with an example of the types of questions and more importantly the types of answers characteristic of the IRB process. You are responsible for the answers, but it may be useful to see examples to see how long they are, and how detailed they need to be.

Following is a form used in one of our recent studies in the lab (Paik, 2011).

Institutional Review Board (IRB)
The Office for Research Protections
205 The 330 Building
University Park, PA 16802 | 814-865-1775 | ORProtections@psu.edu

Submitted by: Jaehyon Paik
Date Submitted: April 09, 2010 10:41:33 AM
IRB#: 33343
PI: Frank E Ritter

Study Title

1> Study Title A New Training Paradigms For Knowledge and Skills Acquisition
2> Type of eSubmission New

Home Department for Study

3> Department where research is being conducted or if a student study, the department overseeing this research study. Industrial and Manufacturing Engineering

Review Level

4> What level of review do you expect this research to need? NOTE: The final determination of the review level will be determined by the IRB Administrative Office. Choose from one of the following: Expedited

5> Expedited Research Categories: Choose one or more of the following categories that apply to your research. You may choose more than one category but your research must meet one of the following categories to be considered for expedited review.
[X] Category 7—Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

Basic Information: Association with Other Studies

6> Is this research study associated with other IRB-approved studies, e.g., this study is an extension study of an ongoing study or this study will use data or tissue from another ongoing study? No

7> Where will this research study take place? Choose all that apply.
[X] University Park

8> Specify the building, and room at University Park where this research study will take place. If not yet known, indicate as such. The research will be held in 319 Information and Science Technology Building.

9> Does this research study involve any of the following centers?
[X] None of these centers are involved in this study

10> Describe the facilities available to conduct the research for the duration of the study. We will mainly use a computer, keyboard, mouse, and joystick to test this study. Through the computer, participants can access the specific website that are developed by us.

11> Is this study being conducted as part of a class requirement? For additional information regarding the difference between a research study and a class requirement, see IRB Guideline IV, "Distinguishing Class-Related Pedagogical (Instructional) Assignments/Projects and Research Projects" located at http://www.research.psu.edu/orp/areas/humans/policies/guide4.asp. No

Personnel

12> Personnel List

PSU User ID	Name	Department Affiliation	Role in this study
jzp137	Paik, Jaehyon	Industrial and Manufacturing Engineering	Co-Investigator
fer2	Frank Ritter	Information Sciences and Technology	Principal Investigator

- **Role in this study** Principal Investigator
  First Name Frank **Middle Name** E **Last Name** Ritter **Credentials** PhD
  PSU User ID fer2 **Email Address** frank-ritter@psu.edu **PSU Employment Status** Employed
  [ ] Person should receive emails about this application
  Mailing Address 316G IST Building
  Address (Line 2)
  Mail Code **City** University Park **State** Pennsylvania **ZIP Code** 16802
  Phone Number 863 3528 **Fax number** **Pager Number** **Alternate Telephone**
  Department Affiliation Information Sciences and Technology
  Identify the procedures/techniques this person will perform (i.e. recruit participants, consent participants, administer the study): This person will administer the whole process of experiments and he will help to recruit participants in his class.
  Describe the person's level of experience in performing the procedures/techniques described above: He has lots of experience doing this kind of experiment. Most of his students who already had a Ph.D. degree did similar experiment from writing an IRB application to doing an experiment.

- **Role in this study** Co-Investigator

First Name Jaehyon **Middle Name** **Last Name** Paik **Credentials**
PSU User ID jzp137 **Email Address** jzp137@psu.edu **PSU Employment Status** Not Employed or Student
[X] Person should receive emails about this application
Mailing Address 125 Washington Place
Address (Line 2)
Mail Code **City** State College **State** Pennsylvania **ZIP Code** 16801
Phone Number 814 876 xxxx **Fax number** **Pager Number** **Alternate Telephone**
Department Affiliation Industrial and Manufacturing Engineering

Identify the procedures/techniques this person will perform (i.e. recruit participants, consent participants, administer the study): This person designed the entire experiments and will perform recruiting participants, receiving consent form from participants, controlling the whole process of experiments, and gathering and analyzing data from participants.

Describe the person's level of experience in performing the procedures/techniques described above: This person is a Ph.D. student in IE department, and he has experience of experiments with human participants in his class. He conducted a similar experiments during his Master student. He also has 5 years in industry, so he has no problem to design and develop the environment.

Funding Source

13> Is this research study funded? Funding could include the sponsor providing drugs or devices for the study. No

NOTE: If the study is funded or funding is pending, submit a copy of the grant proposal or statement of work for review.

14> Does this research study involve prospectively providing treatment or therapy to participants? No

Conflict of Interest

15> Do any of the investigator(s), key personnel, and/or their spouses or dependent children have a financial or business interest(s) as defined by PSU Policy RA20, "Individual Conflict of Interest," associated with this research? NOTE: There is no de minimus in human participant research studies (i.e., all amount must be reported). No

Purpose

16> Provide a description of the research that includes (1) the background, (2) purpose, and (3) a description of how the research will be conducted [methodology: step-by-step process of what participants will be asked to do]. DO NOT COPY AND PASTE THE METHODOLOGY SECTION FROM THE GRANT.

- **Background/Rationale: Briefly provide the background information and rationale for performing the research study.** Most research projects for exploring the effects on learning and retention by varying the training schedule have focused on two type of practice, distributed and massed. The results indicate consistently that the distributed practice has better performance on knowledge and skills acquisition than massed practice. However, a more efficient way might exist, and I assume that a more efficient way is the hybrid practice that uses the distributed practice and massed practice together. Through this study, I will explore more efficient practice strategy.
- **Purpose: Summarize the study's research question(s), aims or objectives [hypothesis].** This study has two objectives, in practical and theoretical way. The first objective is to explore the new paradigm of training strategy for tasks with declarative memory, procedural memory, and, perceptual-motor skill acquisition with different training schedules, such as distributed, hybrid 1 (massed placed in the middle of a regimen), and hybrid 2 (massed placed in the top of a regimen). And the results of each experiment are compared to verify which one is more efficient according to the task type. The second objective is to

verify the results of three types of tasks with the learning and decay theories of the ACT-R cognitive architecture. The ACT-R cognitive architecture provides learning and decay theories to predict human behavior in the ACT-R model. Using these theories, I will explore to verify and summarize the results of the tasks.

- **Research Procedures involving Participants: Summarize the study's procedures by providing a description of how the research will be conducted [i.e., methodology - a step-by-step process of what participants will be asked to do]. Numbering each step is highly recommended. DO NOT COPY & PASTE GRANT APPLICATION IN THIS RESPONSE.** This research follows the order like below: 1. Participants have overall explanation of this research (the objective of the study, which data will be gathered, and so on) 2. After explanation, participants sign a consent form. 3. Participants will have a vocabulary word test for declarative memory, tower of Hanoi game for procedural knowledge, and simple avoiding obstacle game for perceptual motor task game, each game takes no longer 5 minutes. 4. During the task, nothing will be asked to participants. 5. After experiments participants will be asked for not practicing the experiment until their second test.

17> How long will participants be involved in this research study? Include the number of sessions and the duration of each session - consider the total number of minutes, hours, days, months, years, etc. This experiment consists of 8 learning sessions and 1 testing session, and each session takes no longer than 20 minutes. The number of experiment days for participants varies according to the schedule type. Group 1 has 2 days, Group 2 has 8 days, and Group 3 has 4 days for the experiment.

18> Briefly explain how you will have sufficient time to conduct and complete the research within the research period. In the experiment day, Jaehyon will come to the office 1 hour early before the experiment to prepare the experiment, such as turn on the computer, launch the program, and launch a data correction program.

19> List criteria for inclusion of participants: 1. Participants should be older than 18 years2. Participants should have experience using a computer, keyboard, and mouse.

20> List criteria for exclusion of participants: 1. Participants should not have knowledge of Japanese vocabulary.2. Participants should not have any experience of Tower of Hanoi game.

Multi-Center Study

21> Is this a multi-center study (i.e., study will be conducted at other institutions each with its own principal investigator)? No

Participant Numbers

22> Maximum number of participants/samples/records to be enrolled by PSU investigators. NOTE: Enter one number—not a range. This number should include the estimated number that will give consent but not qualify after screening or who will otherwise withdraw and not qualify for inclusion in the final data analysis. This number should be based on a statistical analysis, unless this is a pilot study, and must match the number of participants listed in the consent form. 30

23> Was a statistical/power analysis conducted to determine the adequate sample size? Yes

Age Range of Participants

24> Age range (check all that apply):

[X] 18 - 25 years [X] 26 - 40 years

Participant Information: Participant Categories

25> Choose all categories of participants who will be involved in this research study.
[X] Healthy volunteers

26> Will Penn State students be used as study participants in this research study? Yes

27> Will students be recruited from a Subject Pool? No

28> Will participants be currently enrolled in a course/class of any personnel listed on this application? Yes

29> Describe the steps taken to avoid coercion and undue influence. We will not record any information of participants, so participants could decide to participate without any coercion.

30> Will participants be employees of any personnel listed on this application? No

31> Does this research exclude any particular gender, ethnic or racial group, and/or a person based on sexual identity? No

32> Could some or all participants be vulnerable to coercion or undue influence due to special circumstances (do not include children, decisionally impaired, and prisoners in your answer)? No

Recruitment

33> Describe the specific steps to be used to identify and/or contact prospective participants, records and/or tissue. If applicable, also describe how you have access to lists or records of potential participants. We will recruit participants with two ways. The first way is that participants will be recruited from class (IST 331). We will distribute experiment flyer for participating. The second way is that participants will be recruited by posting and emailing lists in department or college. We will also distribute experiment flyer to the department staffs, and we will ask them to distribute to students. In the experiment flyer, we describe that participants who have knowledge of Japanese vocabulary cannot participate this experiment for screening.

34> Will recruitment materials be used to identify potential participants? Yes

35> Choose the types of recruitment materials that will be used.
[X] Letters/Emails to potential participants [X] Script - Verbal (i.e., telephone, face-to-face, classroom)

36> Describe how potential participants' contact information (i.e., name & address) was obtained. We will ask department staff to broadcast our experiment.

37> Who will approach and/or respond to potential participants during recruitment? Jaehyon Paik

38> Explain how your recruitment methods and intended population will allow you access to the required number of participants needed for this study within the proposed recruitment period. This experiment is not a complex task. It takes no longer than 5 minutes each task, and it also has a simple game that can be attractive to the participants.

39> Before potential participants sign a consent document, are there any screening/eligibility questions that you need to directly ask the individual to determine whether he/she qualifies for enrollment in the study?
[X] Yes

40> During screening/eligibility questions, will identifiable information about these individuals be recorded? No

41> Will investigators access medical charts and/or hospital/clinic databases for recruitment purposes? No

42> Will physicians/clinicians provide identifiable, patient information (e.g., name, telephone number, address) to investigators for recruitment purposes? No

43> Will researchers who are not involved in the care of potential participants review and/or use protected health information before a consent/authorization form is signed in the course of screening/recruiting for this research study (e.g., reviewing medical records in order to determine eligibility)? No

Participant Consent/Assent

44> When and where will participants be approached to obtain informed consent/assent [include the timing of obtaining consent in the response]? If participants could be non-English speaking, illiterate, or have other special circumstances, describe the steps taken to minimize the possibility of coercion and undue influence. The consent form will be given to participants at the first day in the experiment location. Participants should speak and hear English.

45> Who will be responsible for obtaining informed consent/assent from participants? Jaehyon Paik

46> Do the people responsible for obtaining consent/assent speak the same language as the participants? Yes

47> What type of consent/assent will be obtained? Choose all that apply.

[X] Implied consent—participants will not sign consent form (e.g., mail survey, email, on-line survey)

48> One of the following two conditions must be met to allow for a process other than signed informed consent to be utilized. Choose which condition is applicable. Choose only one.

[X] The research presents no more than minimal risk of harm to participants &involves no procedures for which signed consent is normally required outside of the research context.

49> Explain how your study fits into this condition. The experiment that we will have has not any harm for the participants. We just use a computer, mouse, and keyboard, that is, this experiment may part of our life.

50> If multiple groups of participants are being utilized (i.e., teachers, parents, children, people over the age of 18, others), who will and will not sign the consent/assent form? Specify for each group of participants. Participants should read the consent form, and do not need to sign, because we provide implied informed consent form.

51> Participants are to receive a copy of the informed consent form with the IRB approval stamp/statement on it. Describe how participants will receive a copy of the informed consent form to keep for their records. If this is not possible, explain why not. The implied informed form includes contents that "your participation in this research is confidential", and the form will be given to the participants before the experiment.

Cost to Participants: Compensation

52> Will the participant bear any costs which are not part of standard of care? No

53> Will individuals be offered compensation for their participation? No

Data Collection Measures/Instruments

54> Choose any of the following data collection measures/instruments that will be used in this study. Submit all instruments, measures, interview questions, and/or focus group topics/questions for review.
[X] Knowledge/Cognitive Tests

55> Will participants be assigned to groups? Yes

56> Will a control group(s) be used? Yes

57> Choose one of the following:
[X] Other control method

58> Describe the 'other' control method. The difference variable is training schedule in this study.

Drugs/Medical Devices/Other Substances

59> Does this research study involve the use of any of the following? Choose all that apply.

[X] None of the above will be used in this research study

Biological Specimens

60> Will biological specimens (including blood, urine and other human-derived samples) be used in this study? No

Recordings - Audio, Video, Digital, Photographs

61> Will any type of recordings (audio, video or digital) or photographs be made during this study? No

Computer/Internet

62> Will any data collection for this study be conducted on the Internet or via email (e.g. on-line surveys, observations of chat rooms or blogs, on-line interviews surveys via email)? Yes

63> Is there a method in place to authenticate the identity of the participants? No

64> Explain why an authentication method is not in place to identify respondents. We do not collect information of participants.

65> Will data be sent in an encrypted format? No

66> Explain why the data will not be sent in an encrypted format. We do not record information of participants.

67> Will a commercial service provider (i.e., SurveyMonkey, Psych Data, Zoomerang) be used to collect data or for data storage? No

Risks: Potential for and Seriousness of

68> List the potential discomforts and risks (physical, psychological, legal, social, or financial) AND describe the likelihood or seriousness of the discomforts/risks. For studies presenting no more than minimal risk, loss of confidentiality may be the main risk associated with the research. Memorize the Japanese vocabulary may discomfort participants.

69> Describe how the discomforts and risks will be minimized and/or how participants will be protected against potential discomforts/risks throughout the study (e.g., label research data/specimens with code numbers, screening to assure appropriate selection of participants, identify standard of care procedures, sound research design, safety monitoring and reporting). We assume that there is no risk in this experiment. However, if participants feel discomfort in experiment, they can quit the experiment immediately, and they can make a reschedule or they can give up the experiment.

70> Does this research involve greater than minimal risk to the participants? No

Benefits to Participants

71> What are the potential benefits to the individual participants of the proposed research study? (If none, state "None.") NOTE: Compensation cannot be considered a benefit. none.

72> What are the potential benefits to others from the proposed research study? The result may show the needs of new training paradigm.

Deception

73> Does this study involve giving false or misleading information to participants or withholding information from them such that their "informed" consent is in question? No

Confidentiality

74> Describe the provisions made to maintain confidentiality of the data, including medical records and specimens. Choose all that apply.
[X] Locked offices

75> Describe the provisions made to protect the privacy interests of the participants and minimize intrusion. First of all, we do not store any privacy information of participants, and the collected data will be stored in locked office. Only experimenter, Jaehyon Paik, can access the data.

76> Will the study data and/or specimens contain identifiable information? No

77> Who will have access to the study data and/or specimens? Jaehyon Paik (only)

78> Will identifiers be disclosed to a sponsor or collaborators at another institution? No

79> Will a record or list containing a code (i.e., code number, pseudonym) and participants identity be used in this study? No

80> What will happen to the data when the research has been completed? Choose one.
[X] Stored for length of time required by federal regulations/funding source and then destroyed [minimum of 3 years]

81> Is information being collected for this research that could have adverse consequences for participants or damage their financial standing, employability, insurability or reputation? No

82> Will a "Certificate of Confidentiality" be obtained from the federal government? No

HIPAA (Health Insurance Portability and Accountability Act)

83> Will participant's protected health information (PHI) be obtained for this study? No

Radiation

84> Will any participants be asked to undergo a diagnostic radiation procedure while enrolled in this study? No

Physical Activity

85> Will participants be required to engage in or perform any form of physical activity? No

86> Will any type of electrical equipment other than audio headphones be attached to the participants (e.g., EMG, EKG, special glasses)?Submit a letter regarding the most recent safety check of the x-ray equipment being used with the supporting documents for this application. No

Document Upload

ICFS Document 1001 Received 03/22/2010 11:19:22 - Adult Form Revised version of consent form

INSTRUMENTS Document 1001 Received 03/22/2010 11:47:14 - For data collection - All data are recorded in webpage Document 1002 Received 04/09/2010 10:37:36 - The screenshots for the tasks. Document 1003 Received 04/09/2010 10:38:13 - Task2 Document 1004 Received 04/09/2010 10:38:51 - task3

RECRUITMENT Document 1001 Received 03/22/2010 11:20:24 - Recruitment Material Revised version of recruitment mat Document 1002 Received 04/09/2010 10:16:47 - Eligibility Screening This document for Eligibility Scr

SUBMISSION FORMS Document 1001 Received 03/23/2010 09:04:42 AM - Application Auto-generated by eSubmission Approval

• Click ADD to upload a new document for review

• Click REPLACE to upload a revised version of a previously submitted document (the radio button next to the document to be revised must be selected before clicking replace)

• Click REMOVE to delete a document. NOTE: Documents can be deleted at any time prior to submission. If an eSubmission is returned for additional information, only new uploaded documents can be deleted.

• To view a document just click on the document name. The following file types can be uploaded: .doc, .docx, .xls, .xlsx, .ppt, .pptx, .pub, .tif, .tif, .tiff, .txt, .pdf, .rtf, .jpg, .gif

Appendix 7: Considerations When Running a Study Online

Many studies are now moving 'on-line', that is, the subjects are interacting with experiments that are run online through a web browser (http://www.socialpsychology.org/expts.htm provides a list, checked 1/2012). These studies, when properly done, have the possibility to greatly increase your sample size and they certainly have the possibility of providing a much more diverse sample. You can, however, lose experimental control (you won't actually know who is participating in many circumstances), and some technical sophistication may be required to create and use an online study.

Online studies have some special considerations. This section notes a few considerations to keep in mind when running these studies. This section does not consider the choice of tools to run a study, like Amazon's Mechanical Turk or commercial tools to create surveys, because the book focuses on how to start to run studies, not how to design them, implement them, or analyze them, per se. This appendix is also not complete because online surveys is a growing area, and this appendix is deigned to only introduce you to some of the issues in this area. For more complete treatments, see references in the further readings section.

A7.1 Recruiting subjects

If you are recruiting subjects to participate in a study, you might choose to go online to recruit them. If you do so, you should keep in mind that the request should be fair and if your study is under an IRB, how you recruit goes through the IRB as well. We have argued previously (Cheyne & Ritter, 2001) that you should not recruit subjects through unsoliceted direct email, although our university does this at times to distraction. There is a delicate balance here that most people understand how to use in the real world and that we are still learning about in the online world about how to share and draw attention appropriately. Putting the flyer (or announcement) onto a very relevant mailing list can be appropriate if such a mailing list is available and appropriate. Putting announcements of studies up on appropriate web sites can be very appropriate. It can also be appropriate and perhaps overlooked, to put study announcements for online studies out through the channels you would use with a non-online study, such as flyers and class announcements. It seems inappropriate to send announcements about competitions to create 'learning badges' to 'professors at universities we could find', as a private university in Durham, NC recently did.

If your subjects are recruited in a way that you don't see them, you might wish to take a few more demographic measures, depending on your theory and the hypothesis. For example, what country they are in (if your software can't tell from the IP address of their machine), or level of education and first language. One of the clearest summaries of this problem was noted in Lock Haven University's student newspaper (14 October 2010, p. A7) about their online poll "This … poll is not scientific and reflects the opinions of only those Internet users who have chosen to participate. The results cannot be assumed to represent the opinions of Internet users in general, not he public as a whole." If you can work around this restriction, for example, finding best performance or examples, then your results will be worthwhile. If you gather the results as representative, then you are subject to this restriction.

If the link to your software has been widely disseminated, you should have the software fail gracefully after the study is done. For example, if your survey is no longer up on your web server, you could put a page up noting this and thanking those

A7.2 Apparatus

Because the apparatus for gathering the data will be automatic and you will not be able to answer questions that arise (in most cases), the interaction needs to be clear and correct. So, you should run more extensive pilot studies than you would for other studies, examine the interaction experience yourself, and have the PI and other RAs use the apparatus to make sure that there are not typos, wordings that are unclear, or other potential problems. You should also back up information from the server you are using on another machine daily.

If your apparatus is taking timing information you should test this and not take it for granted. It is not that case that a timer that reports the time a user interacted with millisecond precision is generating time stamps that are accurate to a millisecond. This can be difficult to test, but before you report timing data, you should attempt to measure its accuracy.

A7.3 Gaming your apparatus

You should check your data daily. This will be useful to judge if your subject recruitment is going well. It will also be helpful to see if a person or a group is gaming the experiment. They might be doing multiple times because it is fun for them (but this might not provide useful data, or might slow down your server), or they might enjoy 'messing up' your experiment. If you find anomalies, you should contact your PI with these concerns. You should also talk about criteria for removing data that you believe are not provided in earnest.

A7.4 Further readings

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist, 59*(2), 105-117.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., et al. (2011). Comparing the accuracy of RDD telephone surveys and Internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly, 75*, 709-747.

These papers, available online, describe some of the theoretical differences between real world and Internet studies, and online and telephone surveys, including the need to understand who your respondents are.

Joinson, A., McKenna, K., Postmes, T., & Reips, U.-D. (2007). Oxford handbook of Internet psychology. New York, NY: OUP.

This book has a section (8 chapters) on doing research on the Internet.

References

AFB. (2012). Interpreting BLS employment data. http://www.afb.org/Section.asp?SectionID=15&SubTopicID=177.

Al-Harkan, I. M., & Ramadan, M. Z. (2005). Effects of pixel shape and color, and matrix pixel density of Arabic digital typeface on characters' legibility
. *International Journal of Industrial Ergonomics, 35*(7), 652–664.

American Psychological Association. (2001). *The publication manual of the American psychological association (5 ed.)*. New York, NY: American Psychological Association.

Anderson, J. R., Bothell, D., & Douglass, S. (2004). Eye movements do not reflect retrieval processes. *Psychological Science, 15*(4), 225-231.

Avraamides, M., & Ritter, F. E. (2002). Using multidisciplinary expert evaluations to test and improve cognitive model interfaces. In *Proceedings of the 11th Computer Generated Forces Conference*, 553-562, 502-CGF-002. U. of Central Florida: Orlando, FL.

Bethel, C. L., & Murphy, R. M. (2010). Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics, 2*, 347–359.

Boehm, B., & Hansen, W. (2001). The Spiral Model as a tool for evolutionary acquisition. *Crosstalk: The Journal of Defense Software Engineering, 14*(5), 4-11.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments and Computers, 35*, 11-21.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin Company.

Carlson, R. A., & Cassenti, D. N. (2004). Intentional control of event counting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 1235-1251.

Cassavaugh, N. D., & Kramer, A. F. (2009). Transfer of computer-based training to simulated driving in older adults. *Applied Ergnomonics, 40*(943-952).

Cheyne, T., & Ritter, F. E. (2001). Targeting respondents on the Internet successfully and responsibly. *Communications of the ACM, 44*(4), 94-98.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cozby, P. C. (2004). *Methods in behavioral research* (8th ed.). New York, NY: McGraw-Hill.

Crossman, E. R. F. W. (1959). A theory of the acquisition of speed-skill. *Ergonomics, 2*, 153-166.

Darley, J. M., Zanna, M. P., & Roediger, H. L. (Eds.). (2003). *The compleat academic: A practical guide for the beginning social scientist / Edition 2*. Washington, DC: American Psychological Association.

de Groot, A. D., & Gobet, F. (1996). *Perception and memory in chess*. Assen, NL: Van Gorcum.

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9*(1), 1-8.

Dhami, M. K., & Hertwig, R. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959-988.

Digiusto, E. (1994). Equity in authorship: A strategy for assigning credit when publishing. *Social Science & Medicine, 38*(I), 55-58.

Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology*. New York: Dover. (originally published 1885, translated 1913).

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: Bradford Books/MIT Press.

Estes, W. K. (1956). The problem of inference from group data. *Psychological Bulletin, 53*, 134-140.

Fishman, G. A. (2003). When your eyes have a wet nose: The evolution of the use of guide dogs and establishing the seeing eye. *Survey of Ophthalmology, 48*(4), 452-458.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling amplitude of movement. *Journal of Experimental Psychology, 47*(6), 381-391.

Friedrich, M. B. (2008). *Implementierung von schematischen Denkstrategien in einer höheren Programmiersprache: Erweitern und Testen der vorhandenen Resultate durch Erfassen von zusätzlichen Daten und das Erstellen von weiteren Strategien (Implementing diagrammatic reasoning strategies in a high level language: Extending and testing the existing model results by gathering additional data and creating additional strategies).* Faculty of Information Systems and Applied Computer Science, University of Bamberg, Germany.

Hart, S. G., & Staveland, L. E. (1988). Development of the NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-185). Amsterdam: North Holland.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). Repealing the power law: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*, 185-207.

Hill, E. W., & Ponder, P. (1976). *Orientation and mobility techniques: A guide for the practitioner*. New York, NY: American Foundation for the Blind.

Howell, D. C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Thompson Wadsworth.

Jonassen, D. H., & Grabowski, B. L. (1993). *Handbook of individual differences, learning, and instruction*. Hillsdale, NJ: Erlbaum.

Jones, G., Ritter, F. E., & Wood, D. J. (2000). Using a cognitive architecture to examine what develops. *Psychological Science, 11*(2), 93-100.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River: NJ: Prentice Hall/Pearson Education.

Kim, D. S., Emerson, R. W., & Curtis, A. (2009). Drop-off detection with long cane: Effects of different cane techniques on performance. . *Journal of Visual Impairment & Blindness, 103*(9), 519-530.

Kim, J. W. (2008). *Procedural skills: From learning to forgetting.* Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA.

Kim, J. W., Koubek, R. J., & Ritter, F. E. (2007). Investigation of procedural skills degradation from different modalities. In *Proceedings of the 8th International Conference on Cognitive Modeling*, 255-260. Taylor & Francis/Psychology Press: Oxford, UK.

Kim, J. W., & Ritter, F. E. (2007). Automatically recording keystrokes in public clusters with RUI: Issues and sample answers. In *Proceedings of the 29th Annual Cognitive Science Society*, 1787. Cognitive Science Society: Austin, TX.

Kirlik, A. (2010). Brunswikian theory and method as a foundation for simulation-based research on clinical judgment. *Simulation in Healthcare, 5*(5), 255-259.

Kukreja, U., Stevenson, W. E., & Ritter, F. E. (2006). RUI: Recording user input from interfaces under Window and Mac OS X. *Behavior Research Methods, 38*(4), 656-659.

Lintern, G., Sheppard, D. J., Parker, D. L., Yates, K. E., & Nolan, M. D. (1989). Simulator design and instructional features for air-to-ground attack: A transfer study. *Human Factors, 31*, 87-99.

MacWhinney, B., St. James, J., Schunn, C., Li, P., & Schneider, W. (2001). STEP—A system for teaching experimental psychological using E-Prime. *Behavioral Research Methods, Instruments, & Computers, 33*(2), 287-296.

Mané, A. M., & Donchin, E. (1989). The Space Fortress game. *Acta Psychologica, 71*(17–22).

Marron, J. A., & Bailey, I. L. (1982). Visual factors and orientation-mobility performance. *American Journal of Optometry and Physiological Optics, 59*(5), 413-426.
Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57*, 203-220.
Mitchell, M. L., & Jolley, J. M. (2012). *Research design explained* (8 edition ed.). Belmont, CA: Wadsworth Publishing.
Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). New York, NY: John Wiley & Sons.
Moon, J., Bothell, D., & Anderson, J. R. (2011). Using a cognitive model to provide instruction for a dynamic task. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2283-2288. Cognitive Science Society: Austin, TX.
Murphy, R. R., Blitch, J., & Casper, J. (2002). AAAI/RoboCup-2001 Urban Search and Rescue Events: Reality and competition. *AI Magazine, 23*(1), 37-42.
NASA. (1987). *NASA Task Load Index (TLX) V 1.0. Users Manual*: Retrieved 30 November 2004, from http://iac.dtic.mil/hsiac/docs/TLX-UserManual.pdf.
Nerb, J., Spada, H., & Ernst, A. M. (1997). A cognitive model of agents in a commons dilemma. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, 560-565. Erlbaum: Mahwah, NJ.
Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
Nielsen, J. (1994). Usability laboratories. *Behaviour & Information Technology, 13*(1-2), 3-8.
Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of CHI 90*, 249-256. ACM: New York, NY.
Orne, M. T., & Whitehouse, W. G. (2000). Demand characteristics. In A. E. Kazdin (Ed.), *Encyclopedia of psychology* (pp. 469-470). Washington, DC: American Psychological Association and Oxford University Press.
Paik, J. (2011). *A novel training paradigm for knowledge and skills acquisition: Hybrid schedules lead to better learning for some but not all tasks.* Unpublished PhD thesis, Industrial Engineering, Penn State University, University Park, PA.
Pew, R. W., & Mavor, A. S. (Eds.). (2007). *Human-system integration in the system development process: A new look*. Washington, DC: National Academy Press. http://books.nap.edu/catalog.php?record_id=11893.
Ray, W. J. (2003). *Methods: Toward a science of behavior and experience* (7th ed.). Belmont, CA: Wadsworth/Thompson Learning.
Reder, L. M., & Ritter, F. E. (1988). Feeling of knowing and strategy selection for solving arithmetic problems. *Bulletin of the Psychonomic Society, 26*(6), 495-496.
Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not the answer. *Journal of Experimental Psychology : Learning, Memory & Cognition, 18*(3), 435-451.
Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 435-451.
Rempel, D., Willms, K., Anshel, J., Jaschinski, W., & Sheedy, J. (2007). The effects of visual display distance on eye accommodation, head posture, and vision and neck symptoms. *Human Factors, 49*(5), 830-838.
Ritter, F. E. (1989). *The effect of feature frequency on feeling-of-knowing and strategy selection for arithmetic problems.* Unpublished MS thesis, Department of Psychology, Carnegie-Mellon University.
Ritter, F. E., Kukreja, U., & St. Amant, R. (2007). Including a model of visual processing with a cognitive architecture to model a simple teleoperation task. *Journal of Cognitive Engineering and Decision Making, 1*(2), 121-147.

Ritter, F. E., & Larkin, J. H. (1994). Developing process models as summaries of HCI action sequences. *Human-Computer Interaction, 9*, 345-383.

Ritter, F. E., Schoelles, M. J., Quigley, K. S., & Klein, L. C. (2011). Determining the number of model runs: Treating cognitive models as theories by not sampling their behavior. In L. Rothrock & S. Narayanan (Eds.), *Human-in-the-loop simulations: Methods and practice* (pp. 97-116). London: Springer-Verlag.

Roediger, H. (2004). What should they be called? *APS Observer, 17*(4), 46-48.

Rosenbloom, P. S., & Newell, A. (1987). Learning by chunking, a production system model of practice. In D. Klahr, P. Langley & R. Neches (Eds.), *Production system models of learning and development* (pp. 221-286). Cambridge, MA: MIT Press.

Rosson, M. B., & Carroll, J. M. (2002). *Usability engineering: Scenario-based development of human-computer interaction*. San Francisco, CA: Morgan Kaufmann Publishers.

Salvucci, D. D. (2001). An integrated model of eye movements and visual encoding. *Cognitive Systems Research, 1*(4), 201-220.

Salvucci, D. D. (2009). Rapid prototyping and evaluation of in-vehicle interfaces. *ACM Transactions on Computer-Human Interaction, 16*(2), Article 9, 33 pages.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research and Applications Symposium*, 71-78. New York: ACM Press.

Sanderson, P. M., & Fisher, C. A. (1994). Exploratory sequential data analysis: Foundations. *Human-Computer Interaction, 9*(3&4), 251-317.

Schoelles, M. J., & Gray, W. D. (2000). Argus Prime: Modeling emergent microstrategies in a complex simulated task environment. In *Proceedings of the 3rd International Conference on Cognitive Modelling*, 260-270. Universal Press: Veenendaal, The Netherlands.

Schoelles, M. J., & Gray, W. D. (2001). Argus: A suite of tools for research in complex cognition. *Behavior Research Methods, Instruments, & Computers, 33*(2), 130-140.

Seibel, R. (1963). Discrimination reaction time for a 1,023-alternative task. *Journal of Experimental Psychology, 66*(3), 215-226.

Smallman, H. S., & St. John, M. (2005). Naïve Realism: Misplaced faith in the utility of realistic displays. *Ergonomics in Design, 13*(Summer), 6-13.

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). *Psychophysiological recording* (2nd ed.). New York, NY: Oxford University Press.

VanLehn, K. (2007). Getting out of order: Avoiding lesson effects through instruction In F. E. Ritter, J. Nerb, T. O'Shea & E. Lehtinen (Eds.), *In order to learn: How the sequences of topics affect learning* (pp. 169-179). New York, NY: Oxford University Press.

Wagenmakers, E., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science, 17*, 641-642.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Winston, A. S. (1990). Robert Sessions Woodworth and the "Columbia Bible": How the psychological experiment was redefined. *The American Journal of Psychology, 103*(3), 391-401.

Wisconsin DHS. (2006). Sighted guide techniques for assisting the visually impaired. http://www.dhs.wisconsin.gov/blind/adjustment/sightedguide.pdf, Madison, WI.

Woodworth, R. S. (1938). *Experimental psychology*. Oxford, England: Holt.

Index pieces

Index terms from the other ways

OLPC One laptop per child project
Multilingual Fonts study, Multilingual fonts study
Bob, Ying, Edward, Judy

Author index

<include: All authors of papers>

Index terms from the ToC

Index terms from similar books

A

Alternative hypothesis
ANOVA
Apparatus
APA (American Psychological Association)
Authorship

B

Block

C

Cause and effect
Chance for insights
Condition
Conflicts of Interest
Control