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

CogSci 2012: Tutorial Packet
24 July 2012

Prepared by: Jonathan H. Morgan and Frank E. Ritter
The Pennsylvania State University
University Park, PA 16802
jhm5001@psu.edu

Table of Contents

0. Tutorial Overview: Purpose and Schedule

The purpose of this tutorial is to introduce a risk driven approach to experimentation that enables researchers to optimize their research process to meet their research needs in a safe and ethical manner. While all researchers should strive for repeatability in their experimental designs, research questions vary not only in their degree of risk but also in the kinds of risks associated with them. Further, balancing external and internal validity frequently requires identifying those conditions or processes essential for testing the question at hand.

Clinging to ideal types, whether idealized notions of task fidelity or of experimental control, are unhelpful. In the first case, we are likely to be lost in a sea of confounding variables. In the second, our results' applicability is likely to be limited to a rarefied set of conditions; the findings of such a study can also obscure important processes**.** In this tutorial, we describe how to identify potential transfer effects so to achieve design parsimony while also balancing these competing demands. **Thus, we will describe experimental design as a risk driven process that requires us to define our research questions carefully, identify risks early, and revise our risk assessments iteratively.**

In this tutorial, we will walk through the experimental process, with an emphasis on conveying the thought-process involved in running experiments. We provide the book this tutorial draws from[1] to provide you a more in-depth treatment and a practical reference.

Table 1 provides the schedule for the tutorial. Sections 2 through 7 here walk through the major phases of the research process. The organization of this tutorial mirrors that of the book; however, the tutorial provides practical exercises, as well as a greater emphasis on problems confronted by researchers outside of academic settings. Consequently, we will discuss in detail how to streamline your experiments in light of limited access to participants (Section 2), quick ways of converting office space to support experiments (Appendix 2), and the role of simulations in the experimental process (Sections 2 and 3).

Table 1. Tutorial Schedule: Where are we going?

Time	Topic
09:00-09:15	Tutorial Orientation: Who are we, who are you?
09:15-09:45	Risk-Driven Experimental Design
09:45-10:15	Planning Your Experiment
10:15-10:30	Break
10:30-10:45	Anticipating and Addressing Ethical Risks
10:45-11:00	Anticipating and Addressing Risks to Validity
11:00-11:15	Break
11:15-11:45	Running the Experiment: How to Deal with Problems
11:45-12:30	What Happens Afterwards: Debriefing, Analysis, and Reporting

[1] Ritter, F. E., Kim, J. W., Morgan, J. H., & Carlson, R. A. in press, 2013. How to run experiments: A practical guide to research with human participants. Currently 150 pages. Thousand Oaks, CA: Sage.

1. Risk-Driven Experimental Design

In this tutorial and in our book, we advocate a pragmatic approach to the experimental process, summarized in Figure 1. Before planning a study, we suggest first identifying what you hope to learn by conducting your experiment. The experimental process begins first with your experimental question, and your experimental question should reflect both your needs and your goals. There is a purpose behind every question, and matching, explicitly, your question to your purpose is critical. Before rigorously defining your experimental question, picking the methods you intend to use, or identifying the risks associated with those methods, we suggest thinking broadly.

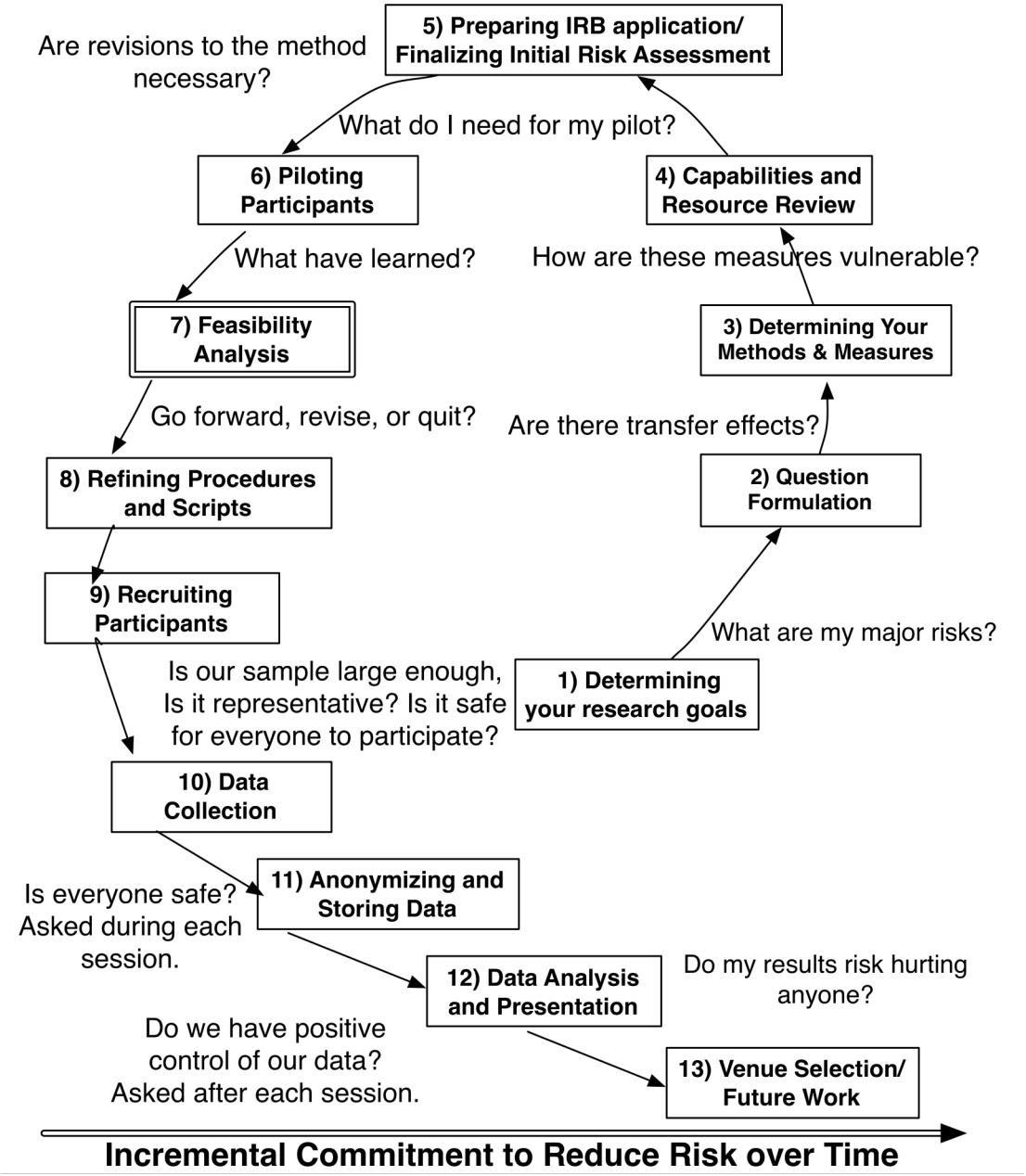

Figure 1.1: Incremental commitment model for experimental design.

The boxes in the figure correspond to decision points. The numbering of the boxes is to provide some sense of the progression of steps over time; however, we acknowledge that are other plausible alternatives. The arrows show the movement between steps, with the spiral illustrating the increasing commitment of resources over time. In this approach, we generally assume that movement from each point to the next represents an increase in cost and thus risk for the total effort. The questions between the boxes are meant to be guiding questions that we discuss more fully in section 2. Finally, the double line border at step seven indicates a pivotal milestone for most projects, when the experimenters must decide whether to commit the significant resources necessary to recruit and run subjects based roughly on the experimental design used during piloting.

1.1. Experimental Questions: Different Questions for Different Purposes

No experimental question is free of risk. When forming a question and choosing the methods by which you will explore it, you are implicitly balancing at least three types of risk*: failure to complete the study, failure to learn anything useful*, and *failure to satisfactorily protect everyone involved*. These risks are directly related to your research question, which in turn directly informs your experimental design. In this instance, economy is a virtue to the extent that we are able to produce findings that meaningfully contribute to what we know, or highlight something important that we do not know, while minimizing these risks. On the other hand, experimental questions should aim to achieve a holistic understanding of a phenomenon.

As noted by Newell (1973) in his famous essay, *You can't play 20 questions with nature and win*, splicing the world into ever more rarefied parts by examining increasing narrow questions does not, in our view, build understanding. With this in mind, we urge you examine research questions with broad implications. In the list below, we provide some guiding questions that we will then use to help us form a research question that meets our research goals while attempting to balance these large sources of risk.

Some guiding questions with regard to experimental designs

1. Are we more interested in determining trends or establishing direct causation?

Question 1 helps us determine if we are simply testing to see if there is a reliable correlation, or if we are interested in identifying a causal relationship. For example, testing for a correlation may be satisfactory for determining whether there is an overall improvement in user performance from one version of an application to the next, especially during early testing and development. Correlation studies can also be useful for determining if users like the new version better. In this situation, controlling the conditions under which users confront the two tools is important for repeatability, asking survey question regarding features is likely to be helpful, and recording the sessions is likely to uncover useful information. On the other hand, it likely that even at this early stage many changes have happened and disambiguating the impact of any one change from the others is less important than verifying that the trend is generally in the right direction. This type of study is relatively easy to complete and provides useful general information; however, a correlation study will not conclusively identify casual relationships nor will it provide more than some indication of the relative importance of

any one factor. **Thus, studies seeking to discover correlations are likely to be informal, draw from a sample of convenience, and generally are not intended for publication – unless the control necessary to establish causal relationships is impossible to achieve with ethical practice.**

For advancing a new theory of design or establishing a definitive explanation for why something works, simply identifying trend data is insufficient. Often, existing theoretical work, correlation studies like the previous example, or anecdotal evidence suggest that there exists a powerful underlying relationship or regularity. Here, disambiguating between factors is crucial and thus more rigorous techniques of experimental control are necessary. Fitts's Law is one such regularity. Fitts's Law predicts the time required to rapidly move to a target is a function of the target's distance and size. Fitts's Law is a powerful regularity operable both in terrestrial and aquatic environments; persistent whether a participant is using his or her hands, feet, or a device; and is almost completely immune to either experimenter or reactive effects. While Fitts's Law is a particularly robust regularity, it highlights an important experimental tradeoff. By moving to a study whose aim is to identify a causal relationship or specify a relationship, we are implicitly positing or at least betting that the relationship in question is robust enough to transfer from an experimental to a real-world setting**.** Thus, we are also betting that the study is worth the additional cost and risk. **Studies attempting to define a relationship or establish causality are generally more formal, require careful consideration of participants and conditions, must address experimenter and participant effects, are likely to require approval by an Institutional Review Board (IRB), and thus are more risky.**

2. Are the dependent measures more or less resistant to environmental pressures or experimenter effects?

Whether attempting to identify a trend or establish a relationship, the response complexity of the phenomenon under study makes isolating the trend or the relationship more or less difficult. At one extreme, the reaction between baking soda and vinegar (famously enshrined in generations of volcano experiments) occurs consistently in the same way every time. We can think of human reaction times as being a bit farther down the continuum. Simple reaction times between people are fairly consistent and largely immune to experimenter effects. On the other hand, distractions can slow reaction times and multiple experimental iterations can lead to learned behaviors that in some instances can effect reaction times. Clinical studies are farther down the continuum still because the response complexity is greater due to the multiple interactions occurring between multiple variables (e.g., the same drug can exhibit a variety of side effects or different degrees of potency across a given sample), and because of socio-cognitive effects, such as the placebo effect, can influence perceived performance.

Finally, deception studies have the potential for a wide array of responses resulting from the experimenter/participant interaction (e.g., experimenters can be poor deceivers), the range of possible participant responses (i.e., the range of emotional responses to a given social situation), and the experiment's dependence on sustained deception across trials and participants (e.g., the potential for participant whistleblowers). We can summarize this point by saying **the greater the potential response complexity the**

greater the need for experimental controls (e.g., double-blind protocols, randomization, etc.), and thus the greater the risk.

3. How important is it to capture the activity/phenomenon as it occurs in situ?

A desire for external validity is not necessarily best served by a full re-creation of the task environment. In general, we support the use of naturalistic studies to better understand users and task environments. For instance, examining mouse activity to develop features to support various special-needs communities by studying users at home or at work makes a great deal of sense. In this case, the activity in question is fairly resistant to situational effects, and in fact those effects, can be considered a tolerable degree of noise associated with studying work processes. Also, there is a clear argument that capturing behavior in situ is important because successfully managing "noise" is part of the task. These studies generally begin by looking for trends that then can be further examined later with greater experimental control if the trends are suggestive.

On the other hand, there are many instances where replicating the full task environment is not helpful because it adds confounding variables. Thus, we must be clear about what we are testing and why. In learning studies and tutoring, we can see where a slavish desire to replicate the task's full complexity can not only be expensive but unhelpful. Learning occurs in stages, particularly for complex tasks where the application of the learned information is multi-dimensional. For instance, there is a clear distinction between knowing that you should perform 30 chest compressions using two fingers when performing Infant CPR as opposed to knowing the proper rhythm or pressure for those compressions. To evaluate where learning is and is not occurring, we must test not simply total task completion but also the participants' knowledge of the component information. Failure to properly perform child CPR could be a failure to know what to do (how many compressions to perform and when to perform them) or it could be a failure to know how to do it (where on the body, at what pressure, and at what speed). Further, failure to respond correctly under simulated emergency conditions could be failure of the participant to learn the information, retain the information, or manage multiple tasks in addition to the CPR task. We are not arguing that cumulative exercises are never useful, only that they are generally better for experts and that they provide a general picture of where learning or failures in learning are occurring. **The more complex the phenomenon being studied the more risk in situ factors can pose to either completing the study or obtaining useful knowledge from it.**

4. Is the activity/phenomenon one specific to a demographic group or population

For many studies, recruiting participants poses a significant risk. There are generally two sources of risk associated with recruiting participants. On one hand, it is often difficult and potentially expensive to recruit a large enough sample, especially outside of academia. A subset of this problem is determining how many participants is enough. With regards to determining the number of participants needed, we recommend either performing a power analysis (Cohen, 1988) or following convention. On the other hand, representativeness can also pose a challenge. While randomization is often cited as the ideal solution to this second problem, recruiting a random sample is frequently difficult and expensive.

Researchers frequently use a multi-level sample to manage costs—they poll 30 organizations and attempt to gather, for example, 10 volunteers from each organization. This is much cheaper than finding 300 random participants. However, the degrees of freedom in the statistical tests must be discounted to acknowledge the partial structuration of the sample pool.

For tasks where we can simulate with some degree of accuracy the processes involved, simulations can help offset this risk, especially when we are interested in identifying a potential trend or discovering a potential z factor.

It may be helpful to again envision a continuum of risk. For some questions, demographic factors have relatively little impact (e.g., the response times on a mouse task). Given a base level of familiarity with this mode of input and sufficient vision and manual dexterity, gender, race, educational background, and age have little impact. We can imagine that these factors might potentially influence a participant's familiarity or ability; however if these factors are held constant, response times are relatively resistant to demographic effects or individual differences. On the other hand, drug trials must carefully consider and control for demographic factors. Gender and age are only two of the demographic factors that can significantly influence a drug's pharmacodynamic profile. Most HCI and HRI studies fall between these two extremes. Nevertheless, we recommend considering participant recruitment early in your planning process. To summarize, **research questions where demographic categories are likely to significantly influence the phenomenon in question are riskier because they generally require larger subject pools and more careful selection.**

5. What are the risks associated with this activity/phenomenon?
    a. Can knowledge obtained through the experimental process harm the participants in any way?
    b. Does the activity/phenomenon itself pose a risk?
    c. Are there cultural or social factors of which we should be aware when investigating this question?

The first four questions are largely design questions, meant to help you determine the size and complexity necessary to explore your research question. In planning your research agenda, we recommend initially modest iterative steps that contribute to a larger framework of knowledge. By iterative, we propose that you periodically reassess whether moving to a more expensive experimental procedure is worthwhile, or if your findings are sufficient to satisfy your current research goals. We also mean iterative in another sense; we encourage you to re-examine your research goals in light of what you have learned. Unanticipated paths can be profitable in research, while a priori notions are frequently dead ends.

With question 5, we are now expanding our notion of risk to include risks to the health and welfare of our experimental team and participants. We encourage you to think broadly. As we generate new types of data and new ways of harvesting and using that data, the risks posed by what we can learn about our participants has grown. The risks posed by biographic or medical data are now fairly well understood. On the other hand, knowledge regarding participant performance in tasks that the participant regards as prestigious or socially significant can also harm people, for example in learning studies,

assessments of professional competency, basic numeracy, or language fluency. In naturalistic studies, information regarding personal web browsing habits can reveal confidential information.

For any study, we encourage you when possible to create codings or procedures that either remove or destroy individual identifiers. For the purposes of experimental validity, we caution you to know what you are aggregating, but in general aggregate your data when possible. We also encourage you to be careful when describing your study, avoid place names, professions (education level often suffices), or references that could connect your participants to a particular time and place. To summarize, **capturing individualized information poses a risk to your participants and entails an ethical duty to all handling the data to safeguard the privacy and confidentiality of the participants to the greatest extent possible.**

6. Do we intend to publish the results of this study?

The publishing goals of a study also influence its design. Published studies require IRB approval, attempt to concretely contribute to our knowledge regarding a phenomenon or regularity, and go through an extensive review process. For all studies, communicating what you have learned to other researchers whether within your company or in a broader community should be a consideration, if only to avoid redundant costs. On the other hand, reporting results also entails costs and should be considered a source of risk. For communicating trend data within your company, a tech report is probably sufficient, whereas papers seeking to specify or generalize a relationship are likely to benefit from external review. **To summarize, reporting is important; however, matching your publication goals and format to your study goals is necessary risk mitigation strategy.**

1.2. Determining What Is Necessary for Transfer of Results

In section 2.1, we introduced some guiding questions for research design. These questions are meant to help you consider your research goals and quickly do an initial risk assessment. In this section, we discuss how to achieve design parsimony by identifying likely transfer effects. This section is specifically oriented towards studies seeking more than probable trend data; but because of ethical considerations, cost constraints, or likely confounding variables, it is important to identify the fundamental dynamics behind a phenomenon of interest and test those dynamics in a safer and more controlled setting. Research examining railroad, automobile, or airline accidents provides good examples of this process.

For example, let us consider the problem of quantifying the impact of texting on driving behavior. For obvious ethical reasons, we cannot simply recruit a random pool of participants, assign them to control and experimental groups (non-texting and texting), and test to see if there is a reliable difference in the number of automobile accidents between the two groups. Instead, we have to use a variety of methods to infer a relationship. To begin with, we can analyze accident reports in a longitudinal retrospective study to see if a correlation between texting and accidents shows up in the data as phones with text capabilities entered the market. This technique can indicate a trend; however, we are unlikely to establish texting's relative influence on the probability

of having an automobile accident compared to other salient factors as many other factors have changed in automobile requirements over that same time-period. Also, other distractions may interact with cell-phone texting. Is texting , for example, any more dangerous than listening to the radio? Listening to the radio and texting are likely to appear together in the data. To disambiguate the relative impact of these two forms of distraction, we need further experimental work.

For subsequent experimental studies to be useful, we must identify what about texting and listening to the radio is different, and then construct an experimental task that allows us to safely test this hypothesis. One difference that comes to mind is that texting requires a shift in the driver's field of vision. In other words, texting forces the driver to look away from the road, while listening to the radio does not. To test this hypothesis, we could turn to a car simulation. In this case, using a simulation is potentially a viable alternative; however, this is not universally true. Consider for a moment NASA's initial studies examining the impact of zero gravity on body functions. For these studies, scientists had to examine the effects of weightlessness through aquatic experiments and by simulating free-fall through rapid descent.

Returning to texting while driving, another strategy is to identify or construct a task that is analogous in its critical points. For instance, we might test the impact of texting on a similar perceptual-motor task, like navigating a go-cart course. Certainly, the controls of a go-cart are simpler. Nevertheless, the control surfaces are fairly similar, and driving a go-cart provides another important similarity (other drivers capable of erratic actions). On the other hand, without significantly modifying the go-car, we have to consider the relative impact of other drivers versus the isolating effect of a car cabin. Is simulating the kind of isolation experienced while driving fundamental to understanding the relative impact of texting vs. listening to the radio? We could conduct further tests to establish this, go ahead and make the necessary modifications to the go-carts, or simply conduct the study using standard go-carts. All these approaches impose risks that the researcher must balance. **The determination of the relative significance of these risks should reflect your what you expect will transfer, in this case what are the essential aspects of texting that are likely to lead to accidents.**

1.3. The Experimental Process

In sections 2.1 and 2.2, we outlined an approach to experimental design. These principles inform the entire experimental process but are most salient to the planning process. In Figure 2.1, we show in yet another way a notional progression of the experimental process. Here, our emphasis is on the relationships between the specific steps as opposed to the risk each entails. Figure 2.1 moves from identifying a research question, planning the study and filing the approval forms with the IRB, setting-up the lab space, conducting pilot experiments, recruiting participants, conducting the study, debriefing the participants, analyzing the results, and reporting the results. Again, we acknowledge that there are alternatives to the ordering of these steps and acknowledge that IRB approval is necessary before proceeding past the pilot stage of most work. We, however, find this ordering useful for introducing the process. Solid arrows indicate step progression while dotted arrows indicate the potential for iterative loops. Figures 2-6

elaborate upon this process. In the following sections, we discuss each of these steps before conducting a practical exercise.

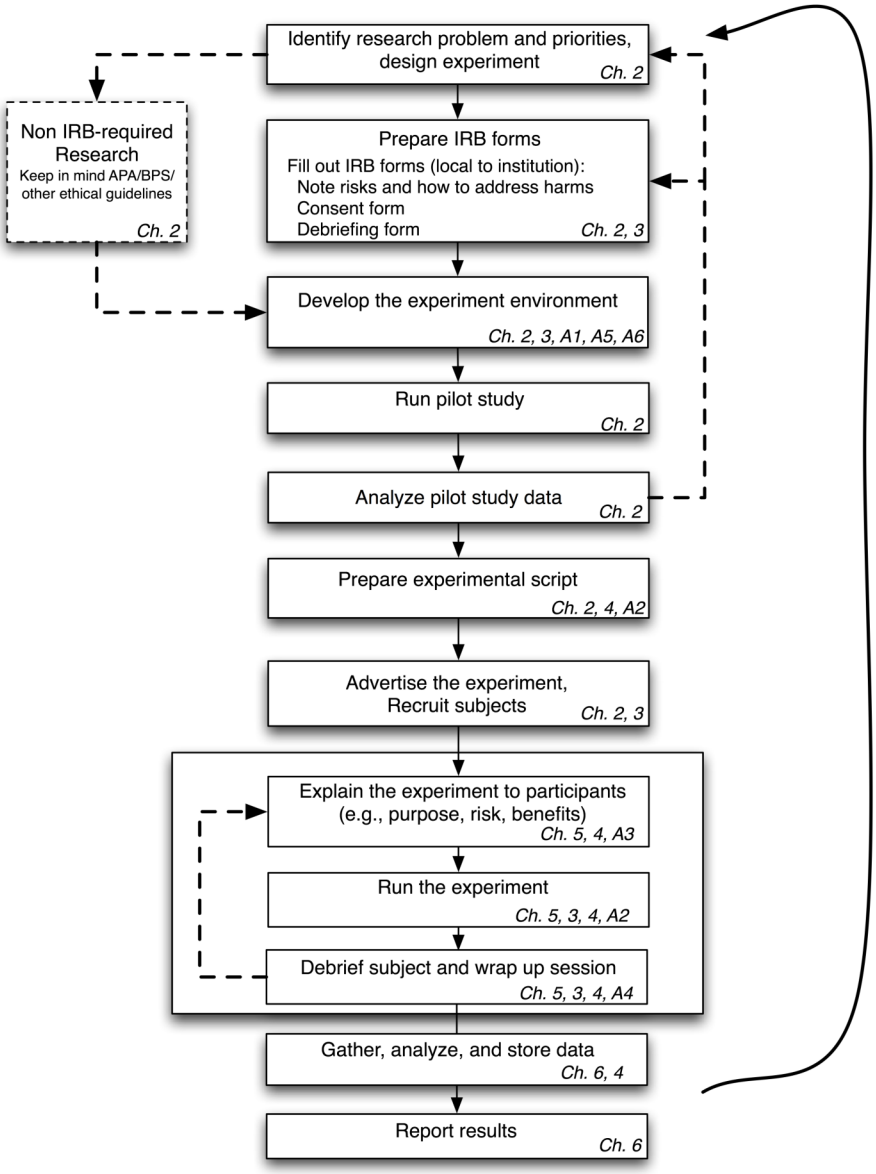

Figure 2.1: An(other) overview of the experimental process[2].

[2] The chapter and section numbers indicated in all these figures correspond to the chapters and sections where we describe these topics in the book.

2. Planning and Setting-Up Your Experiment

Section 2 describes the initial thought process behind planning an experiment (the first box in Figure 2.1), and introduces a framework for weighing risks (Sections 2.2 and 2.3). Here, we will discuss preparing your IRB request, setting-up your experimental environment, and piloting. Remember, we are still defining the task and performing an initial risk assessment. **To segue from the largely conceptual work described in section 2 to the logistical considerations we will cover in this section, we suggest writing your method.** Writing your method at this formative stage is likely to be helpful because this activity will structure your thinking by presenting a general set of considerations. We encourage you to write long in these early drafts, i.e., not only identify the apparatus and procedures you intend to use but describe the setup and steps in considerable detail. Writing long will help in at least two respects: one, you are likely to identify new dependencies introduced by the equipment and experimental environment necessary to execute your method; two, you are more likely to ensure repeatability because you will be leaving less to interpretation. As you begin to formalize your experiment, you are likely to write additional supporting documents, such as an experimental script and checklists.

At this stage, however, we suggest writing your method and then determining if either the equipment, the procedures necessary to use the equipment, or the facilities pose additional risks. For instance, a Human Robot Interaction (HRI) study is likely to involve maintenance, reconfiguring experimental settings across conditions, and general troubleshooting. In a clinical context, preparing different assays is often time consuming, requires technical expertise, and the coordination between clinical technicians and experimenters. These preparatory steps are potential sources of latency or bodily harm, and thus are all risks. You may find after reviewing your available resources and the time commitments associated with either obtaining them or using them that adjustments to your methods are necessary. At this early stage, it may helpful to take advice regarding your assessments. Are they overly optimistic or pessimistic? If optimistic, is there a less resource intensive way to examine your research question? Again, we recommend modest iterative steps followed by re-assessments.

2.1. Summary Discussion

After preparing your method and considering the resources that are likely to be available to you, we suggest preparing your application for IRB approval. For studies where IRB approval is not necessary, we still suggest using the sample IRB form included in the book and in this tutorial (Appendix 1) as a template for completing your risk assessment. We consider this step to be the culmination of your initial risk assessment, started in section 2.1. You may revise your application as you begin to run preliminary tests within your research group to ensure overall feasibility. Nevertheless, preparing your IRB submission or some counterpart will allow you to formalize your risk assessment before potentially investing significant resources or placing anyone at risk.

We highly recommend running pilot studies. In fact, we view them as an essential step in the experimental process. Like other studies, the resources and risks associated with a pilot study should be commensurate with your overall research goals**. In general,**

less formal studies require less formal pilot studies, with the inverse also being true. Conducting pilot studies are a critical risk mitigation strategy and opportunity for refinement. Your pilot study is a model of your experiment; for it to be effective, it must be analogous in all its critical points. Generally, this entails replicating the conditions associated with each experimental condition, recruiting participants that are sufficiently representative of your intended sample, using the intended experimental script, performing each experimental task in the prescribed way, collecting the data from these tasks, and performing preliminary analyses. Pilot studies generally differ from the intended study with respect to the number participants and the number of trials. There are instances where it is necessary to supplement a pilot study with simulations of the task; this is generally true when either obtaining access to the sample population is very difficult, or when the experimental manipulations are very costly or dangerous and thus limited to a few "live" experimental trials (e.g., impact testing for automobiles). In such situations, we would generally recommend exploring less costly alternatives that nevertheless are analogous with respect to the dynamics in question. On the other hand, we recognize that this is not always possible.

Figure 3.1 shows a summary of the experimental process, with boxes indicating steps, arrows sequence, dotted arrows potential iterative loops, and bubbles sub-steps.

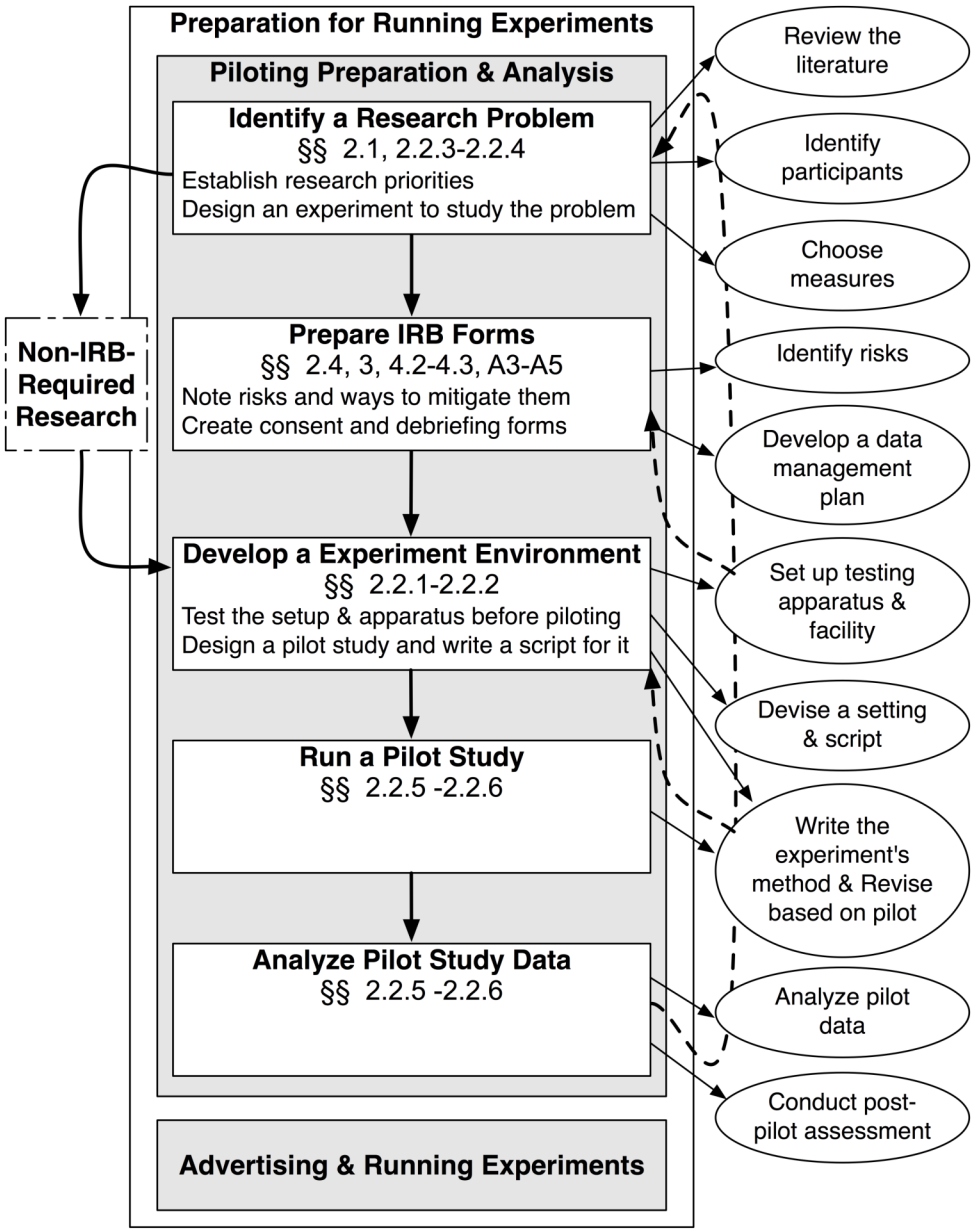

Figure 3.1: A summary of the steps for preparing an experiment.

2.2. Practical Exercise: Description and seed question

This exercise and the following exercises are thought experiments that feature one or two seed questions for you to consider. These seed questions pose different scenarios that we hope help highlight the decision process described in this packet. For this exercise, we would like you to think of an experimental question and outline some potential risks; this process should correspond to the process described in sections 2.1 and 2.2.

Imagine that you are a researcher in a multi-media lab. You are interested in designing a user interface that better enables senior citizens to review, compare, and select health insurance policies. Thus far, your team has identified that zoom functions, audio aids, and generous click and drag functionalities are important features. Your team, however, has not yet tested these features. What features might you add? Also, how would you go about testing the relative efficacy of these features.

3. Anticipating and Addressing Ethical Challenges

In section 2, we described the potential harm posed by information that can be directly connected to a participant or participants. We also noted that you should think broadly when considering ethical risks, including potential social, monetary, or cultural pressures. More generally, 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, 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.

3.1. Summary Discussion

Figure 4.1 notes some major sources of risk and ways to mitigate them. The boxes in this figure indicate risks and where in the book they are described, while the bubbles indicate mitigation strategies for each risk.

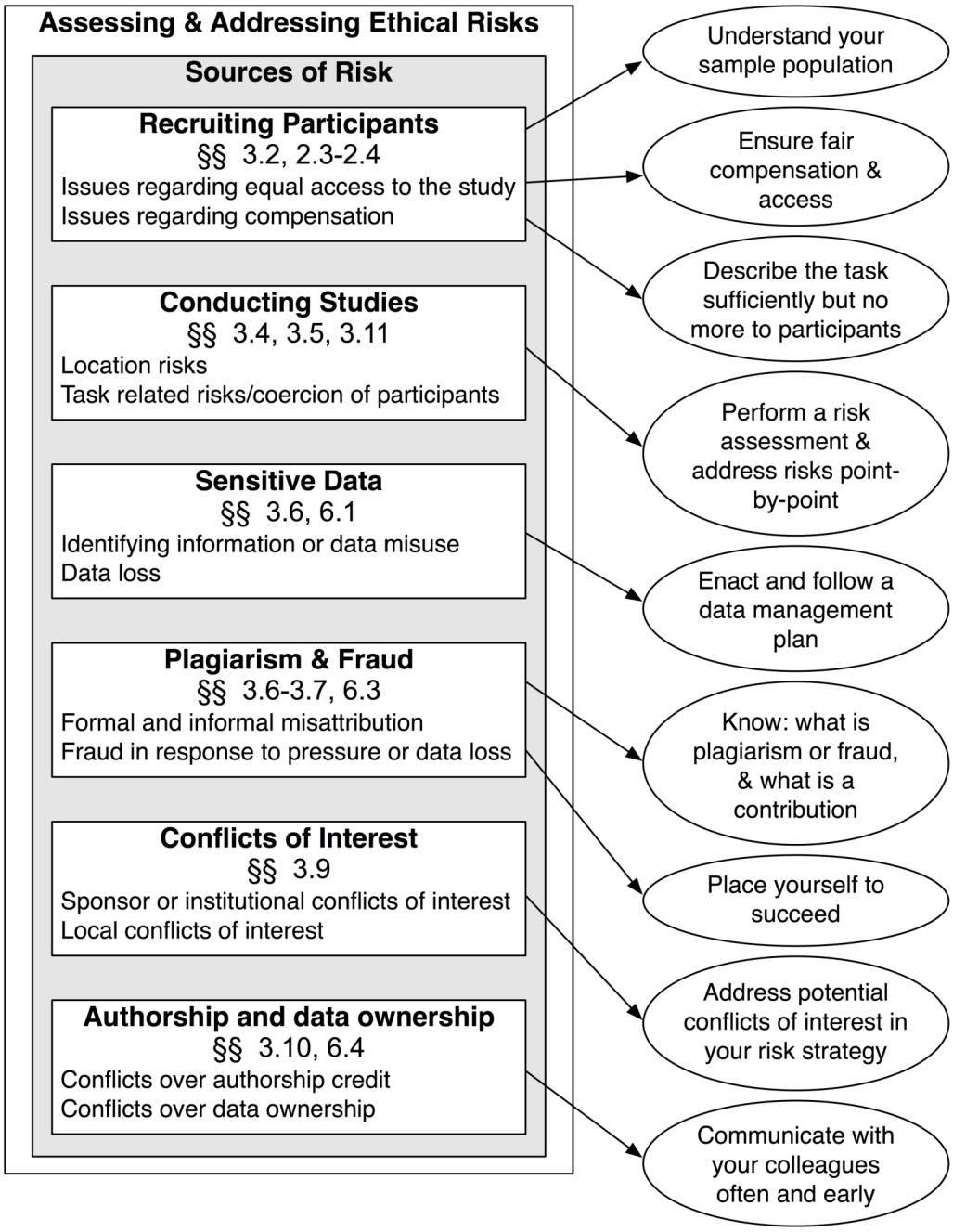

Figure 4. 1: A summary of common ethical risks and associated mitigation strategies.

3.2. Practical Exercise: Description and cases

In this exercise, we present two cases. We do not expect a simple normative answer for either question. Rather, we hope the questions highlight relevant ethical tensions that you may encounter in your own work.

A. In collaboration with the Department of Veterans Affairs, you and your team are evaluating the effectiveness over a period of five years of a learning theory and a tutor based on that learning theory where some of the learners have PTSD. As the study progresses, many of the learners experience significant personal hardship and prolonged unemployment. Does this change in status present an ethical challenge with regards to the participants' freedom of consent? If so, does the veterans' right to participate and their self-felt obligation to help, and their increasing interest in the payments, outweigh this potential threat to consent? Also, what if the nature of the content knowledge (e.g., battlefield first-aid) interacts badly with their PTSD?

B. In screening candidates for a stress study, you discover one of your participant's heart rate suggests a medical condition. (or, in any study situation, a subject arrives in an altered state.) Do you have an ethical obligation to report this to them?

4. Anticipating and Addressing Questions of Validity

Validity refers to the degree to which an experiment leads to an intended conclusion from the data. There are, however, a number of things can reduce the validity of an experiment, and these are known as risks to validity. 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 effects, or the influence of the experimenter's presence on the participants and how this effect can vary across experimenters. 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. Understanding other risks to validity, however, will also help you take steps to minimize biases in your data. 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.

4.1. Summary Discussion

In general, two types of validity, internal validity and external validity, are of interest. Internal validity refers to how well the experimental design explains the outcomes from the experiment. The experimental design includes the independent variables you manipulate, the dependent variables you measure, how subjects are assigned to conditions, and so on. 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. This may be some current event such as news of a terrorist attack or a disaster that influences subjects in a global way leading to better or worse results than would occur at other times. 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: Taking a test can influence scores on 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: Many measuring instruments must be recalibrated regularly. Some instruments need to be recalibrated with changes in humidity. Failure to recalibrate can affect an experiment's results.

Statistical regression: There are risks in 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 participants' normal performance, but scores that are high partly due to chance. On retests, their performance on average will decrease not because of the manipulation but because the 2nd 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 or more boring for subjects, and thus make them less likely to come back 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.

Figure 5.1 provides a summary of the likely sources of risk associated with validity. Boxes indicate sources of risk and where they are discussed in book, while bubbles indicate risk mitigation strategies.

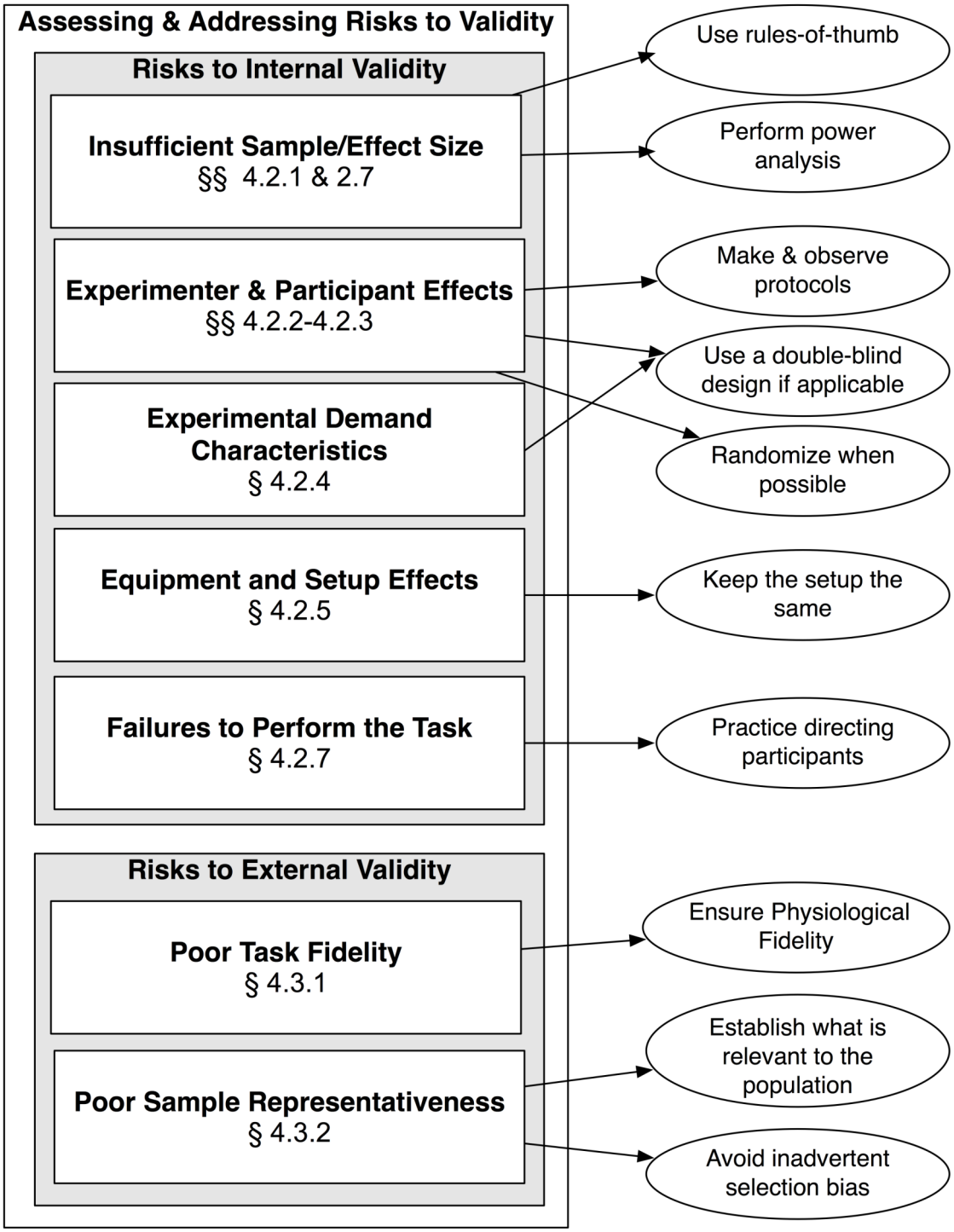

Figure 5.1: A summary of common risks to validity and associated mitigation strategies.

4.2. Practical Exercises: Description and cases

In this exercise, we explore the tension between internal and external validity. While this tension is not universally present, balancing these competing needs is a perennial theme in research.

A. Working with car company, you are testing whether drivers should learn key-based or mouse-based commands for operating a interface being installed in new vehicles. The participants are grouped into 30 person conditions. The company stakeholders are concerned with external validity, and thus want each condition to perform 10 functions deemed mission critical under simulated driving conditions.

The scientists designing the interface want to pursue a more controlled approach, testing the interface in a lab setting using less conditions and more repetitions. The major objection to the lab tests is that they do not simulate either the cognitive load experienced by the drivers or the task switching associated with typical operating conditions.

How would you try to resolve this tension?

5. Running Your Experiment: How to Deal with Problems

As indicated in Figure 6.1 and 6.2, there are many steps associated with running an experiment. In this tutorial we discuss ways of addressing problems during the experimental process. While we will not cover the other steps in great detail here, we do encourage you to refer the book frequently as you prepare to pilot and run your study. We would also add that using a risk driven approach for preparing for and piloting your study will help avoid problems.

On the other hand, 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.

5.1. Summary Discussion

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.

Figure 6.1 shows a notional progression of the preparatory steps for a research session. Boxes represent steps and where they are described in the book, arrows the order of the steps, dotted arrows potential iterative loops, and bubbles sub-steps.

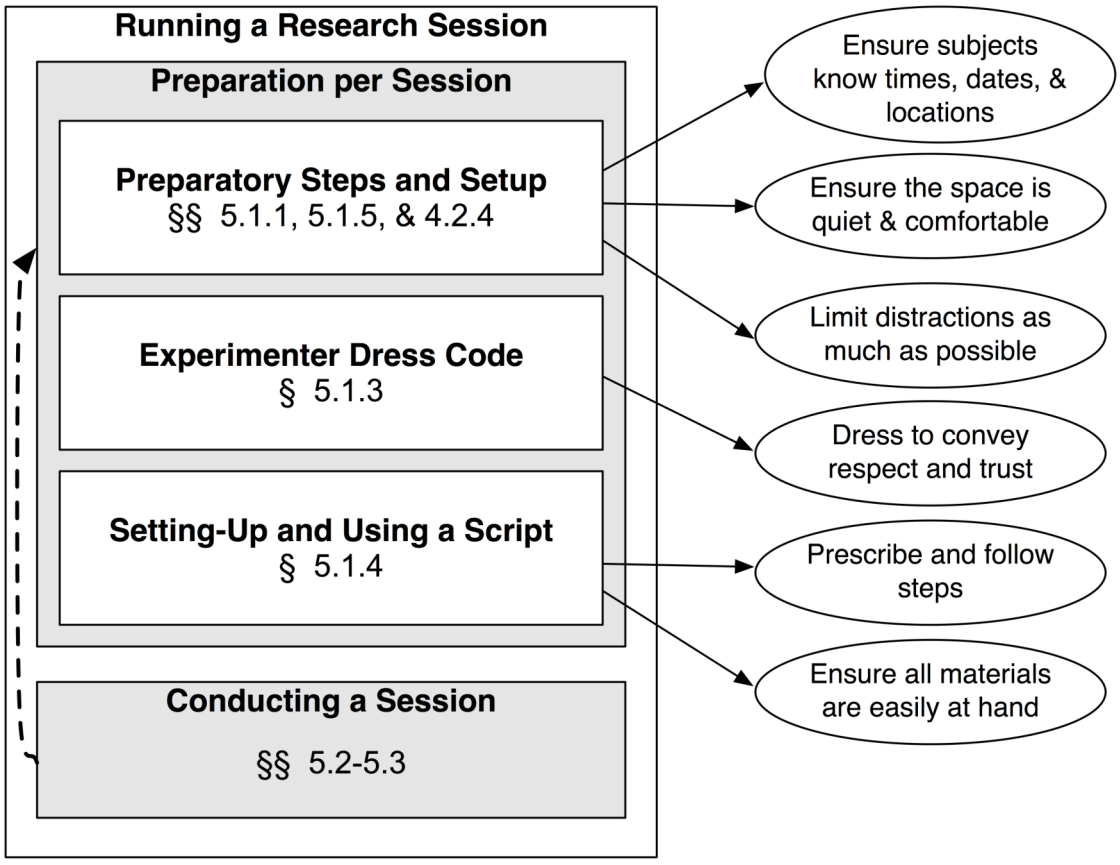

Figure 6.1: Steps for running an experiment.

Moving from preparation to execution, Figure 6.2 shows a notional progression of steps necessary for conducting a research session. Boxes represent steps and where they are described in the book, arrows the order of the steps, dotted arrows potential iterative loops, and bubbles sub-steps.

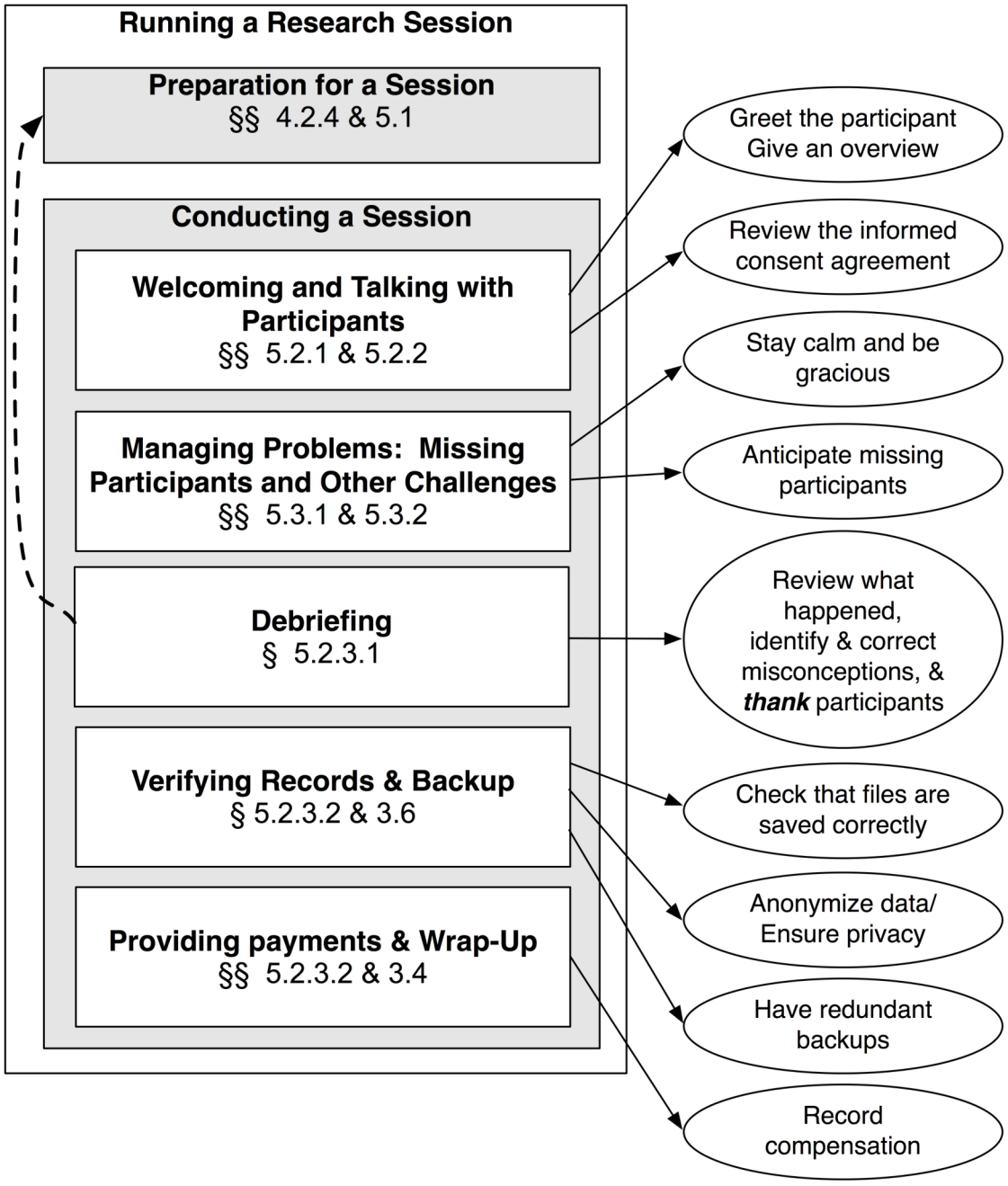

Figure 6.2: Steps for conducting a research session.

5.2. Practical Exercises: Description and cases

In this exercise, we explore potential problems when running a study. Again, these problems do not represent an exhaustive set, but we hope they useful in highlighting some widely applicable steps.

A. In a developmental cognition study, you are working with 10 parents and their infants. You have found in your pilot test that many of the parents are late because the building is confusing. In addition, some of the mothers have inquired whether there might be a play space available for their older children. Right now, you don't have one.

How will instruct your RAs to deal with late parents and older children, particularly children alarmed at being separated from their parents?

B. In a study examining language acquisition in multilingual families (or, indeed any study), you find that some of the participants are concerned about signing the informed consent agreement. While you have provided translations of the agreement, there is still some obvious tension regarding the agreement.

How would resolve this tension?

Also, do you have to exclude participants who are unwilling to sign the informed consent agreement?

6. What Happens Afterwards: Debriefing, Analysis, and Reporting

Concluding a study is important. In this tutorial, we have primarily focused on risks associated with preparing and running a study. Nevertheless, debriefing your participants, analyzing your data, and reporting your results are important considerations on which we will say a few parting words. At the very least, remember to say, "Thank you." Without your participants' cooperation, you will not be successful. Recall that your participants could have used their time differently, and thus deserve to understand why their contribution is significant, what happened to them, and how the data might be used in the future. In other words, **debriefing is an ethical obligation**.

Recalling that the primary purpose behind an experimental investigation is to learn something new, or at least to better understand what we do not know. We would be remiss not to note that ***data loss is a significant risk.*** It may seem redundant, but backup your data often, backup your data on different devices, and backup your data in such a way that you can reference it easily years later. Regarding analyses, we encourage to be open-minded. ***Analyze your data using several different measures, and seriously explore how to best represent your findings by trying multiple formats and techniques***. Do not let poor presentation hinder you from uncovering something important or effectively conveying your results' significance. Figure 7.1 provides a final visual summary of this guidance.

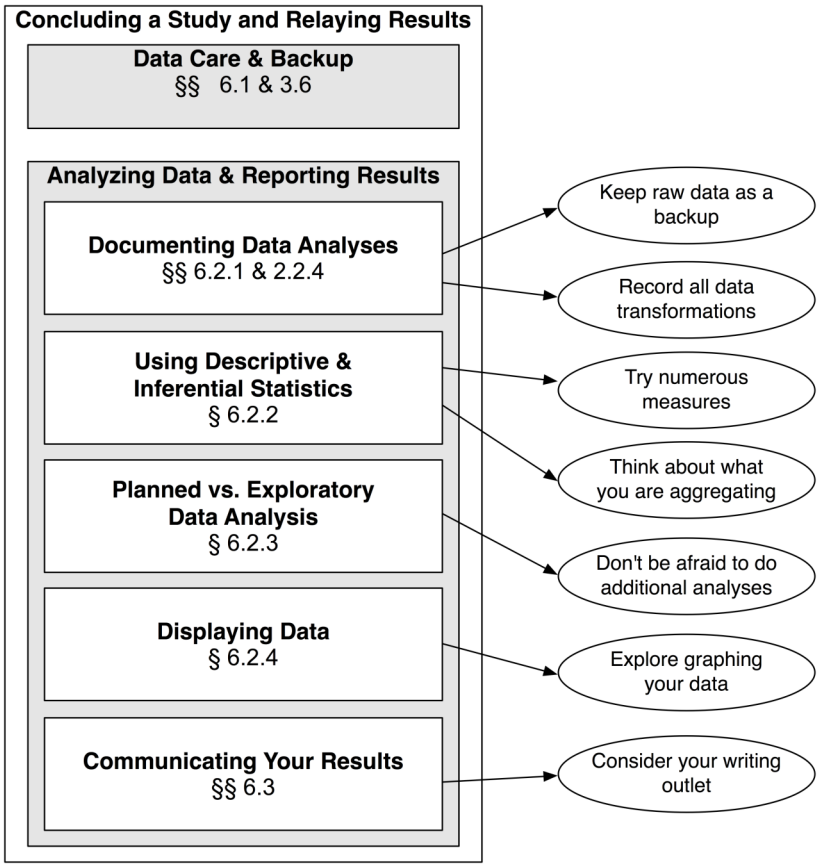

Figure 7.1: Analyzing and relaying results.

6.1. Practical Exercise

In this practical exercise, we ask you to consider publication types, when each is appropriate, what risks are associated with each, and what special considerations might they entail.

Under what conditions would you prepare each of the following publication types: a technical report, a conference paper, or a journal article.

What special considerations do each of these publication types entail, and how do they differ from a book or a thesis?

7. Acknowledgements

We thank Geoffrey P. Morgan for his incisive comments. We also thank Jong Kim and Richard A. Carlson for their continued efforts, parts of this packet draw from material they prepared while writing *How to Run Experiments*…

8. Appendix 1: Example Consent Form (pp. 102-103)

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

9. Appendix 2: Setting-Up Your Lab Space (pp. 74-75)

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 title-of-the-experiment experiment."

Figures 10.1 and 10.2 show two spaces used for running subjects in a psychology department. Figure 10.1 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 10.2 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 10.1: A storage space used as a single-subject data collection station. **Figure 10.2:** An office space used to house multiple data-collection stations.

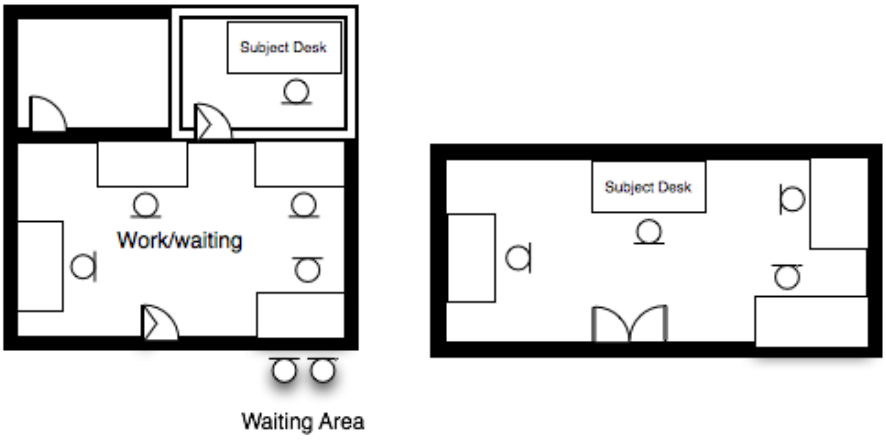

Figure 2-5: Example diagrams of space for running studies. Hollow walls indicate sound proofed walls and a triangle on a door indicates a sweep on the bottom of a drawer to help block sounds.

Draw your space