Experimental Research in HCI

Darren Gergle and Desney S. Tan

Experimental Research in HCI

The experimental method is a technique used to collect data and build scientific knowledge, and it is one of the primary methodologies for a wide range of disciplines from biology to chemistry to physics to zoology, and of course human–computer interaction (HCI).

In this chapter, we learn about the basics of experimental research. We gain an understanding of critical concepts and learn to appreciate the ways in which experiments are uniquely suited to answer questions of causality. We also learn about best practices and what it takes to design, execute, and assess good experimental research for HCI.

A Short Description of Experimental Research

At its heart, experimental research aims to show how the manipulation of one variable of interest has a direct causal influence on another variable of interest (Cook & Campbell, 1979). Consider the research question, "How does the frame rate of a video affect human perception of fluid movement?"

Breaking this down, we can examine several of the elements necessary for good experimental research. The first has to do with the notion of *causality*. Our example question implicitly posits that a change in one variable, in this case frame rate, causes variation in another variable, the perception of fluid movement. More

D. Gergle (✉)
Northwestern University, 2240 Campus Drive, Evanston, IL 60208, USA
e-mail: dgergle@northwestern.edu

D.S. Tan
Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
e-mail: desney@microsoft.com

J.S. Olson and W.A. Kellogg (eds.), *Ways of Knowing in HCI*,
DOI 10.1007/978-1-4939-0378-8_9, © Springer Science+Business Media New York 2014

generally, we often think of two variables, *X* and *Y*; and establishing the notion of causality, which implies that changes in *X* lead to changes in *Y*.

The second thing to note is the idea of *variables*. The researcher needs to manipulate the levels or degree of one or more variables, known as the *independent variables*, while keeping constant other extraneous factors. In this example, our independent variable is frame rate, and we could show the same video at different frame rates, while controlling for other factors such as brightness, screen size, etc. It is also important that we are able to measure the effect that these manipulations have on one or more *dependent variables*. In this case, our dependent variable may be a rating score that captures human perception of fluid movement.

The third thing to note is that our initial question could be formally stated as a *hypothesis* regarding the predicted relationship between frame rate and perception of fluid movement. For example, "An increase in frame rate will increase human perception of fluid movement." The formulation of a hypothesis is important in that it clearly states the parameters of the experiment and communicates the expected relationship. The observed data are then subjected to statistical analysis to provide evidence for or against the hypothesized relationship.

Finally, true experiments require *random assignment* of participants to experimental conditions. Random assignment is critical in establishing equivalent participant groups (with some probability) on both measured and unmeasured characteristics at the outset of the study. This safeguards against systematic biases in assignment of the participants to the experimental conditions, and increases the likelihood that differences across the groups result solely from the treatment to which they are assigned. Without random assignment there exists a risk that attributes of the participants drive the changes in the dependent variable.

Returning to our frame rate example, imagine running a study in which one group of participants watches a video at a low frame rate and a second group watches the same video at a much higher frame rate. You cleverly devise a way to measure perception of fluid movement, recruit participants to come to the lab, and assign the first ten arrivals to the high frame rate condition and the next ten arrivals to the low frame rate condition. After collecting and analyzing your data you find—counter to your hypothesis—that the individuals in the high frame rate condition rated the video as less fluid. Upon further reflection you realize that the participants that showed up first did so because they have a personality type that makes them the kind of person to arrive early. It just so happens that this personality trait is also associated with greater attention to detail and as a result they rate things more critically than the late arrivals. When you do not make use of random assignment, you increase the risk of such confounds occurring.

History, Intellectual Tradition, Evolution

To gain a deeper sensitivity to the role experimental research plays in HCI today, it is helpful to trace its roots, which go back to the development and formalization of the scientific method. Aristotle is often credited in developing initial ideas toward

the search for "universal truths," and the scientific method was popularized and experienced a major emergence with the work of Galileo and others in what is known as the Scientific Revolution of the sixteenth through eighteenth centuries. In a nutshell, scientific inquiry aims to understand basic relations that exist between circumstances and behaviors, with the ultimate goal of aggregating this understanding into a formal body of knowledge.

While experimental research was originally developed as a paradigm for the physical sciences to establish scientific principles and laws, starting in the late nineteenth and early twentieth centuries, psychologists such as Wilhelm Wundt and G. Stanley Hall developed experimental laboratories to investigate human thought and behavior. It quickly became apparent that humans posed a particular challenge for measurement. If humans behaved in a systematic and consistent fashion like the physical world, the application of the scientific method to questions of human behavior would be straightforward. But they do not; individuals vary in their behavior from one moment to the next, and across individuals there can be enormous variability.

As a result of this, researchers in psychology, sociology, cognitive science and information science, as well as the social sciences more broadly, developed new research techniques that were more appropriate for dealing with the vagaries of human behavior in a wide variety of contexts. Most of this early research stayed close to the ideals of the traditional sciences by applying the techniques to support systematic knowledge production and theoretical development regarding human behavior.

As the field of HCI evolved, it became clear that experimental research was useful not only for generating hypothesis-driven knowledge and theoretical advancement but also for informing practical and applied goals. In a recent piece entitled, "Some Whys and Hows of Experiments in Human–Computer Interaction," Hornbæk (2011, pp. 303–305) further argues that experimental research is suitable for investigating process details in interaction as well as infrequent but important events by virtue of the ability to recreate them in a controlled setting. He also highlights the benefits of sidestepping problems with self-reports that stem from faulty human judgments and reflections regarding what lies behind our behaviors and feelings during interaction.

Using an approach known as A/B testing, controlled online experiments are used at large Internet companies such as Google, Microsoft, or Facebook to generate design insights and stimulate innovation (Kohavi, Henne, & Sommerfield, 2007; Kohavi & Longbotham, 2007; Kohavi, Longbotham, & Walker, 2010). Accordingly, some HCI research is more theoretically driven (e.g., Accot & Zhai, 1997; Gergle, Kraut, & Fussell, 2013; Hancock, Landrigan, & Silver, 2007; Wobbrock, Cutrell, Harada, & MacKenzie, 2008), while other research is more engineering-driven with the goal to demonstrate the utility of a technology from a more applied perspective (e.g., Gutwin & Penner, 2002; Harrison, Tan, & Morris, 2010; MacKenzie & Zhang, 1999; Nguyen & Canny, 2005).

Experimental techniques are also widely used in usability testing to help reveal flaws in existing designs or user interfaces. Whether evaluating if one user interface design is better than another; showing how a new recommender system algorithm influences social interaction; or assessing the quality, utility, or excitement

engendered by a new device when we put it to use in the world, good experimental research practices can be applied to make HCI more rigorous, informative and innovative. In fact, many of the benefits of experimental research and its techniques can be seen in HCI studies ranging from tightly controlled laboratory experiments (e.g., MacKenzie & Zhang, 1999; Veinott, Olson, Olson, & Fu, 1999) to "in the wild" field experiments (e.g., Carter, Mankoff, Klemmer, & Matthews, 2008; Cosley, Lam, Albert, Konstan, & Riedl, 2003; Evans & Wobbrock, 2012; Koedinger, Anderson, Hadley, & Mark, 1997; Oulasvirta, 2009).

Advantages of Experimental Research

As a methodology, experimentation has a number of advantages over other HCI research methods. One of the most commonly recognized advantages hinges on its *internal validity*,[1] or the extent to which the experimental approach allows the researcher to minimize biases or systematic error and demonstrate a strong causal connection. When done properly it is one of the few methodologies by which cause and effect can be convincingly established.

In Rosenthal and Rosnow's terms, experimental research focuses on the identification of causal relationships of the form "*X* is responsible for *Y*." This can be contrasted with two other broad classes of methodologies: descriptive studies that aim to capture an accurate representation of what is happening and relational studies that intend to capture the relationship between two variables but not necessarily a causal direction (see Rosenthal & Rosnow, 2008, pp. 21–32).

The experimental method uses precise control of the levels of the independent variable along with random assignment to isolate the effect of the independent variable upon a dependent variable. It also permits the experimenter to build up models of interactions among variables to better understand the differential influence of a variable across a range of others.

It also makes use of quantitative data that can be analyzed using inferential statistics. This allows for statistical and probabilistic statements about the likelihood of seeing the results, and discussion about the size of the effect in a way that is meaningful when comparing to other hypothesized sources of influence.

Experimental research also provides a systematic process to test theoretical propositions and advance theory. A related advantage is that experiments can be replicated and extended by other researchers. Over time, this increases our confidence in the findings and permits the generalization of results across studies, domains, and to wider populations than initially studied. This supports the development of more universal principles and theories that have been examined by a number of independent researchers in a variety of settings.

[1] Much of what makes for good experimental design centers on minimizing what are known as threats to internal validity. Throughout this chapter we address many of these including construct validity, confounds, experimenter biases, selection and dropout biases, and statistical threats.

Limitations of Experimental Research

In general, experimental research requires well-defined, testable hypotheses, and a small set of well-controlled variables. However, this may be difficult to achieve if the outcomes depend on a large number of influential factors or if carefully controlling those factors is impractical. If an important variable is not controlled for, there is a chance that any relationship found could be misattributed.

While an advantage of experimental research is internal validity, the flipside is that these benefits may come at the risk of low *external validity*. External validity is the degree to which the claims of a study hold true for other contexts or settings such as other cultures, different technological configurations, or varying times of the day. A side effect of controlling for external factors is that it can sometimes lead to overly artificial laboratory settings. This increases the risk of observing behavior that is not representative of more ecologically valid settings.

That said, when designing a study there are ways to bolster external validity. Olson and colleagues' paper on group design processes (Olson, Olson, Storrøsten, & Carter, 1993) exemplifies three ways to increase external validity when designing an experiment. First, they chose a task that was a good match for the kinds of activities they had observed in the field—designing an automatic post office—and they tested the task with real software developers to ensure it was an accurate portrayal of everyday work activities. Second, they chose participants for the study that were as close as possible to those they studied in the field. In this case they chose MBA students with at least 5 years of industry experience and who had already worked together on group projects. Third, they assessed the similarity of the behaviors between the laboratory study and their fieldwork on several key measures such as time spent on specific aspects of design and characteristics of the discussions (see Olson et al., 1993, pp. 333–335 and Fig. 4).

Another common challenge for HCI researchers is that they often want to show that their system is "just as good" as another system on some measures while having advantages in other areas. A common mistake is to treat a lack of significance as proof that no difference exists. To effectively establish that things are "just as good" a form of equivalence testing is needed; effect sizes, confidence intervals, and power analysis[2] techniques can be used to show that the effect either does not exist or is so small that it is negligible in any practical sense (for details see Rogers, Howard, & Vessey, 1993).

Furthermore, it should be recognized that hypotheses are never really "proven" in an absolute sense. Instead, we accrue evidence in support of or against a given hypothesis, and over time and repeated investigation support for a position is strengthened. This is critical and points to the importance of replication in experimental work. However, replication is often less valued (and thus harder to publish) in HCI than the novelty of invention. We argue, along with several colleagues, that

[2] G*Power 3 is a specialized software tool for power analysis that has a wide number of features and is free for noncommercial use. It is available at http://www.gpower.hhu.de

as the field matures, replication and extension should become more valued outcomes of HCI research (Wilson, Mackay, Chi, Bernstein, & Nichols, 2012).

Finally, because experimental research is often taught early in educational programs and hence is a familiar tool, it is sometimes force-fit into situations where research questions might have been more appropriately addressed using less formal instantiations of the experimental method or by using other methodologies (for a critique and response, see Lieberman, 2003; Zhai, 2003). A poorly executed experiment may have the veneer of "scientific validity" because of the methodological rigor, but ultimately provides little more than well-measured noise.

How to Do It

In HCI, we often want to compare one design or process to another, decide on the importance of a possible problem or solution, or evaluate a particular technology or social intervention. Each of these challenges can be answered using experimental research. But how do you design an experiment that provides robust findings?

Hypothesis Formulation

Experimental research begins with the development of a statement regarding the predicted relationship between two variables. This is known as a research hypothesis. In general, hypotheses clarify and clearly articulate what it is the researcher is aiming to understand. A hypothesis both *defines the variables involved* and *the relationship between them*, and can take many forms: A causes B; A is larger, faster, or more enjoyable than B; etc.

A good hypothesis has several characteristics. First, the hypothesis should be *precise*. It should clearly state the conditions in the experiment or state the comparison with a control condition. It should also describe the predicted relationship in terms of the measurements used.

Second, the hypothesis should be *meaningful*. One way it can be meaningful is by leading to the development of new knowledge, and in doing so it should relate to existing theories or point toward new theories. Hypotheses in the service of applied contributions can also be meaningful as they reveal something about the design under investigation and can convince us that a new system is more efficient, effective, or entertaining than the current state-of-the-art.

Third, the described relationship needs to be *testable*. You must be able to manipulate the levels of one variable (i.e., the independent variable) and accurately measure the outcome (i.e., the dependent variable). For example, you could be highly influenced by "The Truman Show" (Weir, 1998) and hypothesize that "we are living in a large fish tank being watched by other humans with which we can

have no contact." While the statement may or may not be true, it is not testable and therefore it is speculation and not a scientific hypothesis.

Finally, the predicted relationship must be *falsifiable*. A common example used to demonstrate falsifiability examines the statement, "Other inhabited planets exist in the universe." This is testable as we could send out space probes and show that there are other inhabited planets. However, the lack of detection of inhabited planets cannot falsify the statement. You might argue, "what if every single planet is observed?", but it could be that the detection mechanisms we use are simply not sensitive enough. Therefore, while this statement could be true, and even shown to be true, it is not falsifiable and thus is not an effective scientific hypothesis. You must be able to disprove the statement with empirical data.

Evaluating Your Hypothesis

Once you have established a good hypothesis, you need to demonstrate the degree to which it holds up under experimental scrutiny. Two common approaches for doing this are hypothesis testing and estimation techniques.

Hypothesis Testing

Hypothesis testing, specifically null hypothesis significance testing, is widely used. In the context of HCI, this approach often aims to answer the question "Does it work?" or "Are the groups different?"

The first step in null hypothesis significance testing is to formulate the original research hypothesis as a *null hypothesis* and an *alternative hypothesis*.[3] The null hypothesis (often written as H_0) is set up as a falsifiable statement that predicts no difference between experimental conditions. Returning to our example from the beginning of the chapter, the null hypothesis would read, "Different frame rates *do not* affect human perception of fluid movement." The alternative hypothesis (often written as H_A or H_1) captures departures from the null hypothesis. Continuing with the example, "different frame rates *do* affect human perception of fluid movement."

The second step is to decide on a significance level. This is a prespecified value that defines a tolerance for rejecting the null hypothesis when it is actually true (also known as a Type I error). More formally, this is stated as alpha (α) and it captures the conditional probability, Pr(reject $H_0|H_0$ true). While a somewhat arbitrary choice, the convention of $\alpha = 0.05$ is often used as the threshold for a decision.

The third step is to collect the data (this is a big step that is addressed later in the chapter) and then apply the appropriate statistical test to obtain a p value.

[3] Here we present the Neyman–Pearson approach to hypothesis testing as opposed to Fisher's significance testing approach. Lehmann (1993) details the history and distinctions between these two common approaches.

The p value tells you the probability of obtaining the observed data, or more extreme data, if the null hypothesis were true. More formally, Pr(observed data$|H_0$ true). Therefore, a low p value indicates that the observed results are unlikely if the null hypothesis were true.

The final step compares the observed p value with the previously stated significance level. If $p < \alpha$, then you reject the null hypothesis. Thus, by rejecting the null hypothesis that "Different frame rates do not affect human perception of fluid movement," we bolster the evidence that different frame rates may affect human perception of fluid movement (i.e., we gather additional support for the alternative hypothesis).

While methodologically straightforward to apply, you should recognize concerns with this methodology, so as not to accidentally misinterpret results. These concerns center on its dichotomous "accept" or "reject" outcome, widespread misinterpretation and faulty reporting of results, and inattention to the magnitude of effects and their practical significance (Cohen, 1994; Cumming, 2012, pp. 8–9; Johnson, 1999; Kline, 2004). Several common misunderstandings stem from a misinterpretation of statistical results such as the mistaken belief that a p value indicates the probability of the result occurring because of sampling error or that $p < .05$ means the chances of a Type I error occurring are less than 5 %. Other common mistakes stem from faulty conclusions drawn after accepting or rejecting the null hypothesis such as suggesting that the failure to reject the null hypothesis is proof of its validity, or the common misperception that a smaller p value means a larger effect exists. Finally, researchers should not lose sight of the fact that statistical significance does not imply substantive significance or practical importance. For a detailed description of these and other common mistakes see (Kline, 2013, pp. 95–103).

Estimation Techniques

While the notion of a null hypothesis can be useful to understand the basic logic of the experimental methodology, null hypothesis testing is rarely adequate for what we really want to know about the data. To address some of the challenges of traditional hypothesis testing approaches, contemporary methods rely on *estimation techniques* that focus on establishing the magnitude of an effect through the application of confidence intervals and effect sizes (for recent coverage see Cumming, 2012; Kline, 2013, pp. 29–65).[4] Accessible and thorough descriptions on various estimation techniques can be found in (Cumming, 2012; Cumming & Finch, 2001; Ellis, 2010; Kelley & Preacher, 2012). Bayesian statistics are another alternative that provide greater capability to estimate and compare likelihoods for various hypotheses. For introductions to the Bayesian approach see (Kline, 2013, pp. 289–312; Kruschke, 2010).

[4] We return to effect sizes and confidence intervals in the section "What constitutes good work," where we describe how they can be used to better express the magnitude of an effect and its real world implications.

Estimation techniques retain the notion of a research hypothesis and accruing evidence for or against it, but the emphasis is on quantifying the magnitude of an effect or showing how large or small differences are between groups, technologies, etc. In the context of HCI, estimation approaches aim to answer more sophisticated questions such as, "How well does it work across a range of settings and contexts?" or "What is the size and relative importance of the difference between the groups?" In other words, it aims to quantify the effectiveness of a given intervention or treatment and focuses the analysis on the size of the effect as well as the certainty underlying the claim. This approach may be more appropriate for applied disciplines such as HCI (Carver, 1993) as it shifts the emphasis from statistical significance to the size and likelihood of an effect, which are often the quantities we are more interested in knowing.

Variables

The choice of the right variables can make or break an experiment and it is one of the things that must be carefully tested before running an experiment. This section covers four types of variables: independent, dependent, control variables, and covariates.

Independent Variable

The *independent variable* (*IV*) is manipulated by the researcher, and its conditions are the key factor being examined. It is often referred to as *X*, and it is the presumed cause for changes that occur in the dependent variable, or *Y*.

When choosing an IV, a number of factors should be taken into account. The first is that the researcher can establish *well-controlled variation* in its conditions or levels. This can be accomplished by manipulating the stimuli (e.g., the same movie recorded at different frame rates), instructions (e.g., posing a task as cooperative vs. competitive), or using measured attributes such as individual differences (e.g., selecting participants based on gender or education levels[5]). A group in the condition that receives the manipulation is known as the treatment group, and this group is often compared to a control group that receives no manipulation.

The second is the ability to provide a clear *operational definition* and confirm that your IV has the intended effect on a participant. You need to clearly state how the IV was established so that other researchers could construct the same variable and replicate the work. In some cases, this is straightforward as when testing different input devices (e.g., trackpad vs. mouse). In other cases it is not. For example, if

[5] When using measures such as education level or test performance, you have to be cautious of regression to the mean and be sure that you are not assigning participants to levels of your independent variable based on their scores on the dependent variable or something strongly correlated with the DV (also known as sampling on the dependent variable) (Galton, 1886).

you vary exposure to a warning tone, the operational definition should describe the frequency and intensity of the tone, the duration of the tone, and so on. This can become especially tricky when considering more subjective variables capturing constructs such as emotional state, trustworthiness, etc. A challenge that must be addressed in the operational definition is to avoid an *operational confound*, which occurs when the chosen variable does not match the targeted construct or unintentionally measures or captures something else.

A *manipulation check* should be used to ensure that the manipulation had the desired influence on participants. It is often built into the study or collected at the conclusion. For example, if you were trying to experimentally motivate participants to contribute to a peer-production site such as OpenStreetMap,[6] a manipulation check might assess self-reported motivation at the end of the study in order to validate that your manipulation positively influenced motivation levels. Otherwise the measured behavior could be due to some other variable.

A third important factor to consider is the *range* of the IV (i.e., the difference between the highest and lowest values of the variable). Returning to the example of motivating OpenStreetMap contributions, the range of values you choose is important in determining whether or not motivation levels actually change for your participants. If you gave the "unmotivated" group one dollar, and the "motivated" group two dollars, the difference may not be enough to elicit a difference in cooperative behavior. Perhaps one dollar versus ten dollars may make a difference. It is important that the ranges are realistic and practically meaningful.

Another critical aspect to variable selection is choosing meaningful or interesting variables for your study. In practice this can be even more difficult than addressing the aspects described above. Good variables should be theoretically or practically interesting; they should help to change our way of thinking; they should aim to provide deeper understanding, novel insight, or resolve conflicting views in the literature. Knowing what others have studied and recognizing the gaps in the prior literature can help to achieve this goal.

Dependent Variable

The *dependent variable* (*DV*), often referred to as *Y*, is the outcome measure whose value is predicted to vary based upon the levels of the IV. Common types of dependent variables used in HCI research are self-report measures (e.g., satisfaction with an interface), behavioral measures (e.g., click-through rates or task completion times), and physiological measures (e.g., skin conductance, muscle activity, or eye movements). Picking a good DV is crucial to a successful experiment, and a key element of a good DV is the extent to which it can accurately and consistently capture the effect you are interested in measuring.

Reliability is important when choosing a DV. A measure is perfectly reliable if you get the same result every time you repeat the measurement under identical

[6] http://www.openstreetmap.org

conditions. There are many steps that help to increase the reliability of a DV[7] and decrease the variability that occurs due to measurement error. For each of your DVs, try to:

- *Clearly specify the rules for quantifying your measurement*: Similar to the construction of the IV, you need to be able to detail exactly how your DV was constructed and recorded. This includes formulating coding and scoring rules for the quantification of your measure, or detailing the calculations used when recording the value of your DV. If you cannot clearly articulate your rules you will likely introduce noise into your measure.
- *Clearly define the scope and boundaries of what you are going to measure*. You need to articulate the situations, contexts, and constraints under which you collect your data. For example, suppose you want to measure online content sharing by counting how many times in a session people perform link sharing to external web content. What counts as "a session?" What counts for "link sharing?" Does it have to be original content or can it be a copy of someone else's post? Does it have to be the actual link to a URL or could it be a snippet of content?

Validity is another important consideration when choosing your DV. It is not enough to know that a measure is reliable. It is also important to know that a measure captures the construct it is supposed to measure—if it does so it is considered a valid measure. The following lists ways to assess the validity of your measures, in order from weakest to strongest[8]:

- *Face validity* is the weakest form of validity. It simply means that your measure appears to measure what it is supposed to measure. For example, imagine you propose to measure online satisfaction with a web purchasing process by counting the number of positive emoticons that are present in the purchase comments. You feel that the more a person uses positive emoticons, the more satisfied they were, so "on its face" it is a valid measure.
- *Concurrent validity* uses more than one measure for the same construct and then demonstrates a correlation between the two measures at the same point in time. The most common way to examine concurrent validity is to compare your DV with a gold-standard measure or benchmark. However, concurrent validity can suffer from the fact that the secondary variable or benchmark for comparison may have the same inaccuracies as the DV under investigation.
- *Predictive validity* is a validation approach where the DV is shown to accurately predict some other conceptually related variable later in time. The prototypical example is the use of high-school GPA to predict first year's GPA in undergraduate classes.

[7] When developing new measures it is important to assess and report their reliability. This can be done using a variety of test–retest assessments.

[8] Sara Kiesler and Jonathon Cummings provided this structured way to think about dependent variables and assessing forms of reliability and validity.

- Best practice is to make use of *standardized* or *published* measures when available.[9] The major benefit is that a previously validated and published measure has been through a rigorous evaluation. However, the challenge in using preexisting measures is to make sure that they accurately capture the construct you want to measure.

The *range* of the DV is another important aspect to consider. A task that is so easy that everyone gets everything correct exhibits a "ceiling effect"; while a task so difficult that nobody gets anything correct exhibits a "floor effect." These effects limit the variability of measured outcomes, and as a result the researcher may falsely conclude there is no influence of the IV on the DV.

Related to range is the *sensitivity* of the dependent variable. The measure must be sensitive enough to detect differences at an appropriate level of granularity. For example, an eye tracker with an accuracy of 2° will not be able to capture a potentially meaningful and consistent difference of ½°.

The final thing to consider when selecting a DV is *practicality*. Some data are more accessible than others and therefore are more viable for a given study. Some practical aspects to consider: How often do the events occur? Will the cost of collecting the data be prohibitive? Can you access all of the data? Will your presence influence the behavior under observation?

Control Variable

In addition to independent and dependent variables, there are a number of potential variables that must remain constant; otherwise you run the risk of fluctuations in an unmeasured variable masking the effect of the independent variable on the dependent variable. A *control variable* is a potential IV that is held constant. For example, when running reaction time studies you need to control lighting, temperature, and noise levels and ensure that they are constant across participants. Holding these variables constant is the best way to minimize their effects on the dependent variable. Unlike an independent variable, a control variable is not meant to vary but rather stay constant in order to "control" for its influence on the DV. For any given experiment there are an infinite number of external variables, so researchers make use of theory, prior literature and good discretion to choose which variables to control.

Covariate

While a good experiment does its best to control for other factors that might influence the dependent variable, it is not always possible to do so for all extraneous

[9] It should be noted that numerous surveys and questionnaires published in the HCI literature were not validated or did not make use of validated measures. While there is still some benefit to consistency in measurement, it is less clear in these cases that the measures validly capture the stated construct.

variables. *Covariates* (or, somewhat confusingly, "control variables" in the regression sense) are additional variables that may influence the value of the dependent variable but that are not controlled by the researcher and therefore are allowed to naturally vary. These are often participant baseline measures or demographic variables for which there is theoretical rationale or prior evidence suggesting a correlation to the dependent variable. The idea is that they need to be controlled because random assignment is not perfect, particularly in small samples, and therefore experimental groups may not have been completely equivalent before the treatment. When this is the case, covariates can be used to control for potential confounds and can be included in the analysis as statistical controls.

Research Designs

Up to this point we have discussed the basic components of experimentation. In this section we examine various research designs that bring together these components in ways to best accrue evidence for a research hypothesis. While there are several texts that provide extensive coverage of experimental designs, we focus on designs most commonly used in HCI research. We examine randomized experiments (also known as "true experiments") and quasi-experiments and discuss the differences between the two designs.

Randomized Experiments

We begin by examining a class of experiments known as randomized experiments (Fisher, 1925). Their distinguishing feature is that participants are *randomly assigned* to conditions, as this results in groups that, on average, are similar to one another (Shadish, Cook, & Campbell, 2002, p. 13). In order to keep from conflating attributes of the participants with the variables under investigation, randomized, unbiased assignment of participants to the various experimental conditions is required for all of these study designs. This can often be done through a coin toss, use of a table of random numbers, or a random number generator.[10]

We begin by describing single-factor designs that allow us to answer questions about the relationship between a single IV and a single DV. We then move on to examine more advanced designs for multiple IVs and a single DV (known as *factorial designs*) as well briefly discuss those designs involving multiple IVs and multiple DVs.

[10] Lazar and colleagues (Lazar, Feng, & Hochheiser, 2010, pp. 28–30) provide a step-by-step discussion of how to use a random number table to assign participants to conditions in various experimental designs. In addition, numerous online resources exist to generate tables for random assignment to experimental conditions (e.g., http://www.graphpad.com/quickcalcs/randomize1.cfm).

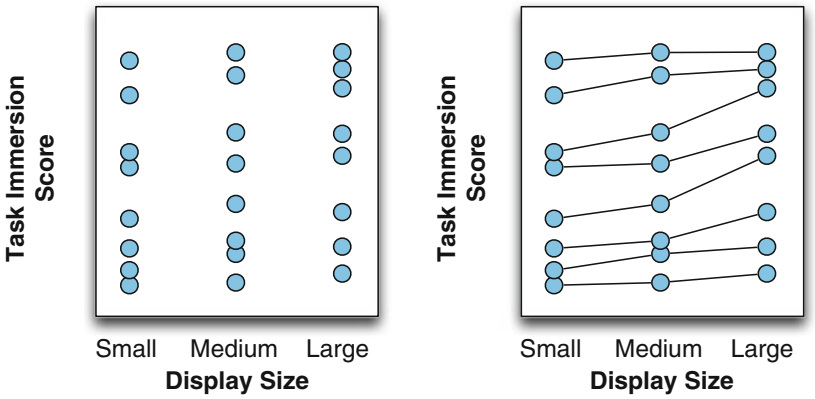

Fig. 1 Example demonstrating the ability to more easily detect differences with within-subjects design (*right*) as compared to a between-subjects design (*left*) when there are large individual differences in participants' scores

Between-Subjects Design

The *between-subjects design* is one of the most commonly used experimental designs and is considered by many to be the "gold standard" of randomized experimental research. Participants are randomly assigned to a single condition (also known as a level of the IV).

Consider, as an example, a rather simple research question that aims to assess the effect that *display size has on task immersion*. Your independent variable is display size, and it has three conditions: small, medium, and large. You also have a single dependent variable: a behavioral measure of task immersion. Let us also assume that you have 24 participants enrolled in the study. In a between-subjects design, you would assign eight participants to the small display size condition, eight to the medium display size condition, and the remaining eight to the large display size condition.

Most of the benefits of a between-subjects design derive from the fact that each participant is only exposed to a single condition. As a result, there is no concern that the participant will learn something from their exposure to one condition that will influence measurement of another condition. This is particularly useful for scenarios where the participant may learn or develop competencies that could affect their performance in another condition.

If fatigue is likely to be an issue, between-subjects designs have the advantage of shorter duration because the subjects are only exposed to a single experimental condition. Between-subjects designs also afford lengthier experimental tasks for the same reason.

However, there are also a number of drawbacks to the between-subjects design. The biggest disadvantage occurs when there are large individual differences in performance as measured by the DV. This can translate into a failure to detect a difference when there is one (i.e., a Type II error) because the higher individual variance makes it difficult (relatively speaking) to achieve a statistically significant result. Figure 1 demonstrates this difference. Looking at the data in the left-hand panel

from 24 different individuals (eight in each of the conditions) you would be hard pressed to suggest there is a difference in performance across the groups. However, consider the same spread of data drawn from eight individuals who participated in each condition as shown in the right-hand panel (this is a within-subjects design and is discussed in the next section). In this figure each individual's data points are connected by a line and it is easy to see that in all cases the score increases, even though there is a great deal of variability across participants in their baseline levels of task immersion.

Another disadvantage is that the groups of participants assigned to the various conditions may not be equivalent and may systematically vary along an unseen dimension—and this is why random assignment is a crucial requirement of all true experiments. In addition, there are a number of practical challenges with between-subjects designs such as the need for a larger number of participants to examine an equivalent number of experimental conditions.

Within-Subjects Design

A within-subjects design is one in which participants are assigned to all conditions (i.e., all levels of the IV) or have repeated exposure to a single condition (known as a repeated measures design). Returning to our research question regarding display size and task immersion, each of the 24 participants would be exposed to the small, medium, and large display sizes.

The main advantage of within-subjects designs stems from the fact that the same participant is examined under numerous conditions, which effectively allows them to serve as their own control. When there is a large amount of individual variation, a within-subjects design is a more sensitive design for capturing differences across conditions because you can look at differences within a person experiencing the conditions. If everyone, independent of level of performance, is better on one than the other, then you can still find significant differences. The general rule of thumb is that when there are large individual differences with respect to the dependent variable, a within-subjects design will be more effective.

Within-subjects designs can also be highly efficient. The number of participants required to show a significant difference among experimental conditions is reduced compared to a between-subjects design. For example, if you have three conditions, you would need three times the number of participants in a between-subjects design as you would in a within-subjects design. In factorial designs, which we discuss later, the multiplier can be even greater. This efficiency can be particularly helpful when studying populations that are high-risk, rare (e.g., participants with rare disabilities or in an isolated locale) or difficult to recruit in large numbers or for long periods of time (e.g., celebrities, high-level executives, and medical surgeons).

The major disadvantage to within-subjects design is that once participants are exposed to a condition they may be altered in a way that will impact their behavior in other conditions. For example, if a participant learns something in the first exposure that influences their performance, there is no way to have them "unlearn" what

Table 1 Summary table for choosing a between-subjects design or a within-subjects design

Choose…	
Between-subjects design	Within-subjects design
• When there are small individual differences, but large expected differences across conditions	• When there are large individual differences (i.e., high variance across participants with respect to the dependent variable(s) of interest)
• When learning and carryover effects are likely to influence performance	• When tasks are unlikely to be affected by learning and carryover effects are unlikely to occur
• When fatigue may be an issue	• When working with rare or hard to reach populations

was just gained. This is particularly problematic with studies that involve learning or insight solutions where you suddenly understand something that was previously perplexing. More generally, these problems are known as *order effects*, since the results may be influenced by the order in which participants go through the conditions.

Another challenge for within-subjects designs has to do with fatigue. For tasks that are physically or cognitively challenging, having the subject perform several repeated tasks is not an ideal solution. If participants become tired, the data can be influenced by the fatigue. Spreading the testing out over time (e.g., hours or days) can resolve the fatigue issue but can introduce unwanted extraneous influences, not to mention the practical issues of researcher time and scheduling.

Learning and fatigue are issues that often come up in HCI research. For example, consider a study examining information retrieval in two different websites. If the participants learn about the basic structure of the website in the first trial, they will carry over this knowledge to the same task on the second site. These types of problems are more generally known as *carryover effects*, and there are several ways to minimize their impact that are described in the following sections. For a summary of factors to consider when choosing between a between-subjects design and a within-subjects design, see Table 1.

Counterbalancing. Counterbalancing helps minimize carryover and order effects by controlling the presentation order of conditions across participants so that each condition appears in each time period an equal number of times. In our display size study this means we would want the small, medium, and large display size conditions to appear in each presentation position an equal number of times.

Complete counterbalancing requires that the participants are balanced across all possible treatment orders. In a simple experiment with few conditions, this is relatively easy. Table 2 shows our three-level experiment with its six possible orderings. However, as the number of conditions increases, the potential orderings grow at a rate of $n!$, where n is the number of conditions.

Since complete counterbalancing is only feasible for small numbers of conditions—with only five conditions there are 120 different orderings needed—researchers have developed a compromise approach where each treatment occurs equally often in each position. *Latin square designs*[11] (Cochran & Cox, 1957;

[11] There are numerous online resources for obtaining Latin square tables (e.g., http://statpages.org/latinsq.html).

Table 2 Complete counterbalancing for a 3-level IV (A,B,C), within-subjects experiment

Participant	First treatment	Second treatment	Third treatment
1	A (small display)	B (medium display)	C (large display)
2	A	C	B
3	B	A	C
4	B	C	A
5	C	A	B
6	C	B	A

Table 3 A Latin square design for a 4-level IV (A,B,C,D), within-subjects experiment

Participant	First treatment	Second treatment	Third treatment	Fourth treatment
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

Fisher & Yates, 1953; Kirk, 1982; Rosenthal & Rosnow, 2008, pp. 192–193) are a form of *partial counterbalancing* that ensure that each condition appears in each position an equal number of times. Table 3 presents a simple Latin square for four conditions.

A common question that arises regarding Latin square designs is what to do with the next cluster of participants. One option would be to continue to use the same Latin square over and over again for each new cluster of participants (e.g., 1–4, 5–8, 9–12, and so on). If using this approach, be sure to test whether the partial counterbalancing is systematically related to the effects of the conditions. An alternative is to generate new Latin squares for each additional cluster of participants. This has the advantage of reducing the likelihood that the partial counterbalancing correlates with the results, but the disadvantage is that this correlation cannot be tested in a straightforward way (for details on these approaches see Kirk, 2013, Chaps. 14–16).

Even better than standard Latin square designs are *balanced Latin square designs* where each condition precedes and follows each other condition equally often. This can help to minimize sequential effects.[12] For example, in Table 3 notice that A precedes B in three of the four rows. A better design can be seen in Table 4 where A precedes B an equal number of times as B precedes A. A balanced Latin square (Bradley, 1958; Williams, 1949) can be constructed for an even number of conditions using the following algorithm for the first row of the square: 1, 2, n, 3,

[12] This approach only balances for what are known as first-order sequential effects. There are still a number of ways in which repeated measurement can be systematically affected such as nonlinear or asymmetric transfer effects. See (Kirk, 2013, Chap. 14) or other literature on Latin square or combinatorial designs for more details.

Table 4 A balanced Latin square design for a 4-level IV (A,B,C,D), within-subjects experiment

Participant	First treatment	Second treatment	Third treatment	Fourth treatment
1	A	B	D	C
2	B	C	A	D
3	C	D	B	A
4	D	A	C	B

$n-1$, 4, $n-2$, …, where n = the number of conditions. Each subsequent row is constructed by adding 1 to the value of the preceding row (or subtracting 1 if the value is equal to n).[13]

Factorial Designs

Up to this point we have focused on experiments that examine a single independent variable at a time. However, in many studies you will want to observe multiple independent variables at the same time, such as gender, display size and task complexity. In such a design each variable is called a *factor*, and the designs that make use of many factors are *factorial designs*.[14] Factorial designs can be either between-subjects, within-subjects, or both in what is known as *mixed factorial designs*[15] (or split-plot designs).

The number of factors and their conditions can be multiplied to yield the total number of conditions you will have for a given experiment. A study with two factors, each with two conditions would yield four total conditions. The name for such a design would be a 2×2 factorial. There is no theoretical limit to the number of factors that can be included in a study; however, there are practical limitations since each additional factor can drastically increase the number of participants needed and the analysis and interpretation become correspondingly complex. For example, a $3 \times 3 \times 4 \times 2$ design would yield 72 different configurations that would each require enough participants to have a well-powered experiment. If you were using a between-subjects design and including 10 participants in each condition, you would need 720 participants! If you used a mixed factorial or within-subjects design you could reduce the overall number of participants needed, but you would have to be careful about fatigue, ordering and carryover effects.

[13] If your experiment has an odd number of conditions, then two balanced Latin squares are needed. The first square is generated using the same method described in the text, and the second square is a reversal of the first square.

[14] As a side note, Latin square designs are a within-subject version of a general class of designs known as fractional factorial designs. Fractional factorial designs are useful when you want to explore numerous factors at once but do not have the capacity to run hundreds or thousands of participants to cover the complete factorial (see Collins, Dziak, & Li, 2009).

[15] In practice, mixed factorial designs are often used when examining different groups of participants (e.g., demographics, skills). For example, if you are interested in differences in user experience across three different age groups, a between-subjects factor may be age group (teen, adult, elderly), while a within-subjects factor may be three different interaction styles.

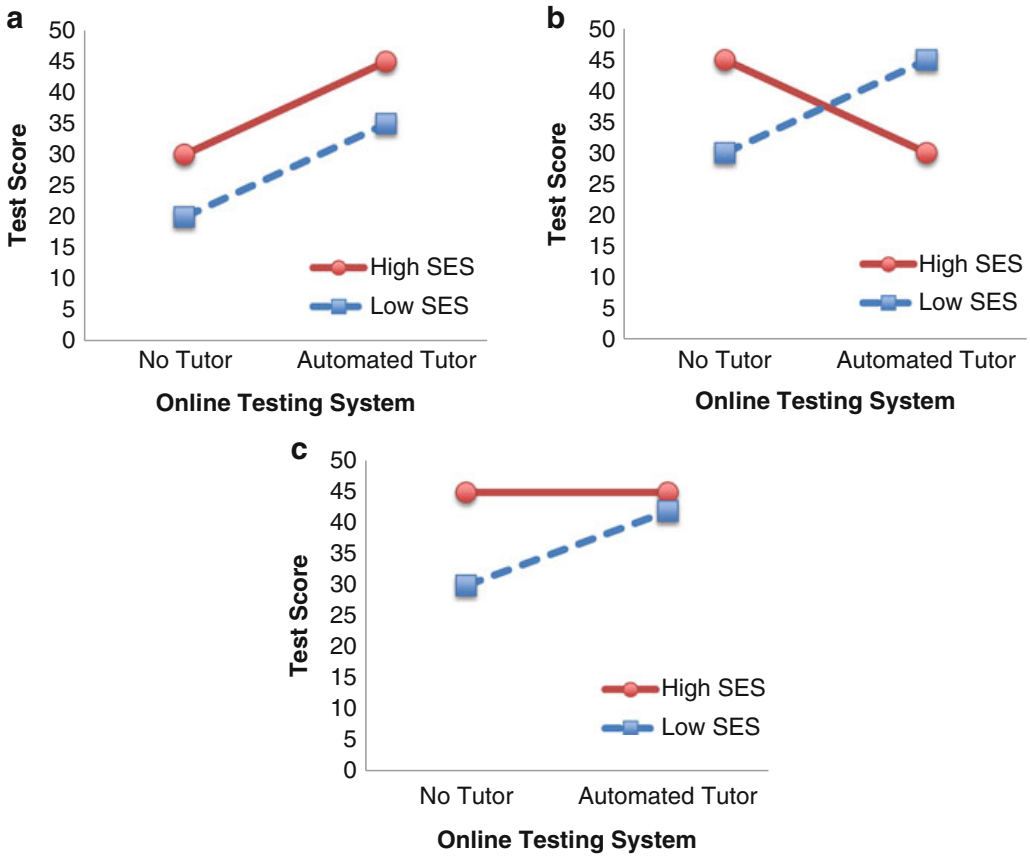

Fig. 2 Three sample outcomes from a 2×2 factorial design showing (**a**) two main effects, no interaction, (**b**) no main effects but a crossover interaction, (**c**) two main effects and an interaction (Color figure online)

Main effects and interactions. A major strength of factorial designs is that they allow you to build up a more complex understanding of the simultaneous relationship between several independent variables and the dependent variable. In other words, you can examine both *main effects* and *interactions*. A main effect is the influence of a single independent variable upon the dependent variable. An interaction occurs when the effect of one independent variable on the dependent variable varies according to the levels of another independent variable.

Figure 2 illustrates a subset of the possible outcomes for a 2×2 factorial design that examines test performance for two different groups—low socioeconomic status (SES) and high SES—using one of two different online testing systems (one with an automated tutor and one without). For this design there are two potential main effects: SES and online testing system. There is also an SES×online testing system interaction.

Figure 2a shows what a graph might look like with a main effect of SES where high SES scores higher than low SES (i.e., the red line is higher than the blue line) and a main effect of online testing system where the automated tutor scores higher than the no tutor system (i.e., the average of the two points on the left is lower than the average of the two points on the right).

Figure 2b exemplifies another possibility and shows why investigating interactions[16] can be helpful. If you only examined the main effects (by averaging across levels of the second IV) you would come to the conclusion that there is no difference between the groups or systems tested. However, there is a clear interaction. This form of interaction, known as a crossover interaction, shows that the effect on the dependent variable goes in opposite directions for the levels of the variable under investigation—and it can mask differences at the main effect level.[17]

Figure 2c shows a result that suggests that both the online tutoring system and SES may matter. However, there is an SES × online testing system interaction that reveals the automated tutoring system primarily benefits the low SES group.

Determining Sample Size and Statistical Power

When designing an experimental study it is important to plan for the number of participants needed. The use of too many participants can be a waste of time and money, and it runs the risk of uncovering small or even meaningless differences. Too few participants, and you may fail to detect differences that actually exist. Ideally you want an estimate that will allow you to reach a conclusion that is accurate with sufficient confidence.

A systematic approach to determining sample size depends on the particular experimental design, number of conditions, desired level of statistical confidence ($p < .05$ is often used), desired sensitivity or power to detect differences (80 % power is often used), a good estimate of the variability in the measurements, and an understanding of what a meaningful difference is in the context of your experiment.

Bausell and Li (2002) and Cohen (1988) provide excellent coverage of the topic, and Kenny (1987, Chap. 13) provides a nice example for studies with a small number of experimental conditions. There are also numerous web resources for determining appropriate sample sizes such as http://www.statsoft.com/textbook/power-analysis/. Most statistical software packages also provide tools to generate visual representations called power curves that can be particularly useful when you are less confident of your measurement estimates.

Quasi-Experimental Designs

In HCI research true random assignment may be impractical, infeasible, or unethical. For example, consider a study that compares performance in a classroom with a new technological innovation versus a traditional classroom without it. In this case, the students are not randomly assigned but instead are preselected based on the classroom to which they were previously assigned. When this is the case, there is a

[16] Note that common transformations of the data (e.g., logarithmic or reciprocal transformations) can affect the detection and interpretation of interactions. Such transformations are performed when the data deviate from the distributional requirements of statistical tests, and researchers need to be cautious when interpreting the results of transformed data.

[17] For factorial designs with more factors, higher-order interactions can mask lower-order effects.

risk that other factors may come into play when measuring the dependent variable. For instance, the teacher of the technological innovation class may also be a better teacher and that may be the primary reason for a performance enhancement.

Quasi-experimental designs[18] aim to address the internal validity threats that come about from a lack of randomization. The designs tend to vary along two primary dimensions: those with or without control or comparison groups; and those with or without pre- and post-intervention measures.

Non-equivalent Groups Design

The non-equivalent groups design is one of the most commonly applied quasi-experimental designs in HCI. The goal is to measure changes in performance that result from some intervention. However, this design lacks the random assignment of participants to experimental groups. This is why it is called "non-equivalent" groups—because the two groups are not equivalent in a way that they would be if random assignment had been used. In many ways it is structured like a typical pre-test/post-test design with a control or comparison group:

Group A: Obs_1–[Intervention]– Obs_2
Control Group: Obs_1 Obs_2

The ideal outcome from such a design is that there is little difference in the pre-intervention measure (pre-test) but large differences in the post-test measure. In other words, the more likely that the groups are equivalent at pre-test time (Obs_1), the more confidence we can have in the differences that appear post intervention (Obs_2). However, there are still a number of threats to internal validity. One is that there are latent attributes of Group A that are not revealed in the pre-testing but that interact with the intervention in some way. Another is that the groups are receiving uneven exposure over time between the pre-test and post-test. Returning to the classroom example, if the teacher in the classroom with the technological innovation also exposes students to something else related to the dependent variable, then we run the risk of misattributing the changes in the dependent variable.

Interrupted Time-Series Design

The interrupted time-series is another popular quasi-experimental design.[19] It infers the effects of an independent variable by comparing multiple measures obtained

[18] For more detailed coverage of quasi-experimental designs see (Cook & Campbell, 1979; Shadish et al., 2002).

[19] Time-series approaches have particular statistical concerns that must be addressed when analyzing the data. In particular, they often produce data points that exhibit various forms of autocorrelation, whereas many statistical analyses require that the data points are independent. There are numerous books and manuscripts on the proper treatment of time-series data, many of which reside in the domain of econometrics (Gujarati, 1995, pp. 707–754; Kennedy, 1998, pp. 263–287).

before and after an intervention takes place. It is often used when there is a naturally occurring event that takes place or in field studies where it is infeasible to have a control group.

The basic form of an interrupted time-series design relies on a series of measurements with knowledge of when an intervention, treatment or event occurred, followed by another series of measurements:

Group A: Obs_1–Obs_2–Obs_3–[Intervention]–Obs_4–Obs_5–Obs_6

If the intervening event or treatment had an effect, then the subsequent series of observed values should experience a quantifiable discontinuity from the preceding measurements. While the easiest change to see is an immediate shift from a flat line, there are numerous ways in which the changes can manifest including intercept or slope changes.[20]

However, there are some major threats to internal validity that must be assessed with time-series designs in HCI. The primary concern hinges on whether another influential event took place at the same time as the intervention (e.g., a major press release about your online news system broke at the same time you implemented a new algorithm aiming to improve online contributions), or whether there was significant mortality or drop out that occurred between the first set of measures and the second (e.g., the participants that were not contributing much dropped out completely for the later stages of the study).

Strengthening Causal Inferences from Quasi-Experimental Designs

For both non-equivalent groups and interrupted time-series designs, there are a number of concerns that arise regarding internal validity, most of which result from the lack of random assignment or use of a control group. To address these concerns, a number of variations have been developed.

The first integrates *treatment removal* into the design.[21] If the intervention is reversible, then the research design can include this to bolster the causal evidence. The first part of the study is the same as the interrupted time-series design, but the second half includes a removal of treatment followed by additional measures:

Group A: Obs_1–Obs_2 [+Intervention] Obs_3–Obs_4 [−Intervention] Obs_5–Obs_6

Naturally, you can extend this design to have *multiple additions and deletions*. If the dependent variable is sensitive to the intervention you should see it respond to each addition and deletion of the treatment, increasing the likelihood that you have identified a causal effect.

---

[20] For a detailed discussion of interrupted time-series designs see (Shadish et al., 2002, pp. 171–206).

[21] These are also known as A-B-A or withdrawal designs, and are similar to many approaches used for small-N or single-subject studies with multiple baselines. For further details see (Shadish et al., 2002, pp. 188–190).

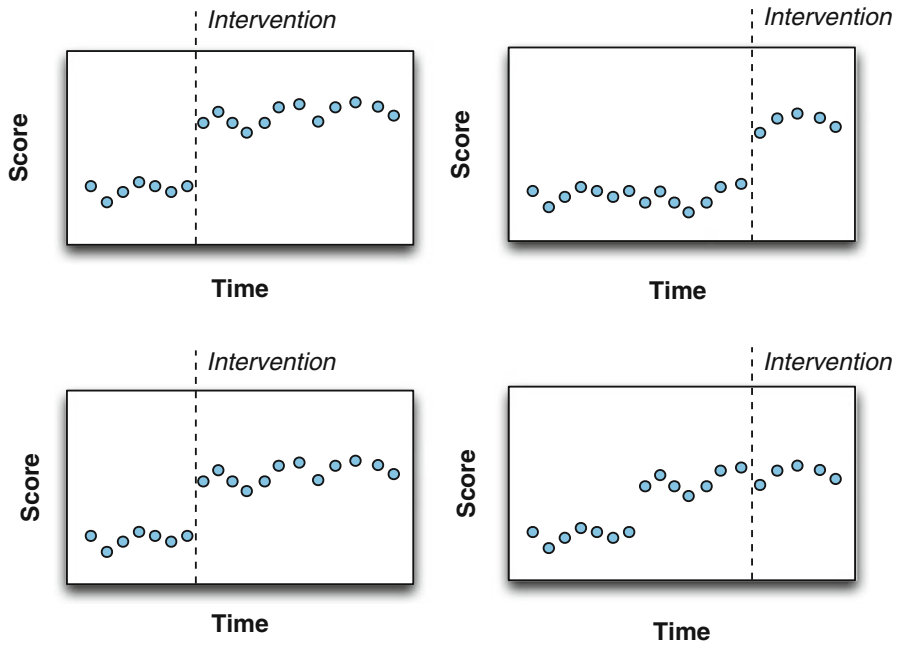

Fig. 3 An illustration of the benefit of time-series with switching replications for detecting or minimizing the potential influence of exogenous factors. The *top* two figures illustrate a discontinuity in the time-series that occurs inline with the intervention, while in the *bottom* two figures the discontinuity in the data occurs at the same time point regardless of the intervention (i.e., it is more likely due to an exogenous influence)

A second variation uses *switching replications* (Shadish et al., 2002, pp. 146–147). Switching replications make use of more than one group in order to introduce interventions at different times:

Group A: Obs_1–[Intervention]–Obs_2–Obs_3–Obs_4–Obs_5–Obs_6
Group B: Obs_1–Obs_2–Obs_3–[Intervention]–Obs_4–Obs_5–Obs_6
Group C: Obs_1–Obs_2–Obs_3–Obs_4–Obs_5–[Intervention]–Obs_6

If the treatment truly causes a shift in the dependent variable, then you should see the shift whenever the intervention takes place (see top panel of Fig. 3), whereas if the change in the dependent variable was caused by another external factor (e.g., the aforementioned press release), then the shift would occur at the same time regardless of the timing of the intervention (see bottom panel of Fig. 3). Introducing the intervention at different times helps to counter internal validity arguments regarding the influence of simultaneous events, history, or even mortality issues.

Finally, you can couple the approaches of interrupted time-series and non-equivalent control group designs. This design can offer some of the strongest support for causal inferences:

Group A: Obs_1–[Intervention]–Obs_2–Obs_3–Obs_4–Obs_5–Obs_6
Group B: Obs_1–Obs_2–Obs_3–[Intervention]–Obs_4–Obs_5–Obs_6
Group C: Obs_1–Obs_2–Obs_3–Obs_4–Obs_5–[Intervention]–Obs_6
Control Group: Obs_1–Obs_2–Obs_3–Obs_4–Obs_5–Obs_6

In summary, there are several advantages to quasi-experimental designs. One of the biggest is that they permit research investigations that may not be possible using randomized experimental approaches. For HCI researchers, this often includes cases where the investigation best takes place in a naturalistic context. To demonstrate an effect in its natural environment is a convincing argument regarding its real-world significance, and demonstrates that even with all of the external factors that may come into play in a natural setting, the effect still has an influence. In this way, quasi-experimental designs can be particularly well suited to the evaluation of contextual social issues, evaluations in educational settings, or for use with hard to reach or limited populations as well as in many assistive technology environments.

The major disadvantage of quasi-experimental designs is the threat to internal validity. In this section we have discussed several ways in which to address validity concerns. However, you may not know of the problem until it is too late. Another more practical challenge is that these designs, when done properly, often require the use of additional participants to serve as controls and comparison groups. If you are working with a limited population this can be challenging. Finally, these designs can be more complex to implement and also to analyze.

Statistical Analysis

Just as important as the research design is planning the statistical analysis ahead of time in a way that ensures you can draw the appropriate conclusions from your experiments. Once the data have been collected, descriptive and inferential statistical analysis methods are used to assess confidence in the findings. A detailed treatment of statistics is beyond the scope of this chapter and the reader is instead directed to the references at the end of the chapter.

Over the years, however, we have found that having a pointer of where to look for the right statistical tests is just as important both when designing an experiment and when evaluating the results of a study. There are numerous flow charts available online for choosing the right statistical test for a given experimental design (e.g., http://abacus.bates.edu/~ganderso/biology/resources/stats_flow_chart_v2003.pdf).

What Constitutes Good Work?

So what ultimately constitutes good experimental research? As Robert Abelson describes in his seminal book, "Statistics as Principled Argument," it's M.A.G.I.C. Abelson (1995) suggests that a persuasive argument using experimental results relies upon the Magnitude, Articulation, Generality, Interestingness, and Credibility of your research. While the MAGIC acronym was originally developed to describe data analysis and its presentation, it can also be useful when thinking about what constitutes good experimental research.

The MAGIC Criteria

Magnitude. The magnitude of your research has to do with understanding the size of the effect being reported and whether it is big enough to have "real world" implications. Assessing magnitude requires more than just obtaining a *statistically significant* difference between experimental conditions. In fact, as previously discussed, a common mistake is to report the *p* value as indicative of an effect's magnitude. The *p* value, critically, depends on two things: the size of the difference between the two groups[22] and the size of the sample. Thus, you can achieve a significant result with a small sample when there is a really big difference between your groups; alternatively, you can also achieve a significant result with very small differences between the groups, if you have a large enough sample. As a result, a better (i.e., smaller) *p* value does not mean it is a "more significant" or "bigger" effect. Reporting *p* values will tell you if there is a significant difference between the groups under investigation; it will not in and of itself tell you whether the difference is meaningful.

The concept of *effect size* can help to determine whether the difference is meaningful. Effect sizes are used to quantify the size of the mean difference between groups (Abelson, 1995, pp. 45–52; Cohen, 1988; Grissom & Kim, 2005; Rosenthal & Rosnow, 2008, pp. 55–58). They can be reported either in original units (i.e., the raw score) or in standardized forms, the latter of which can also be used when the variable's units do not have an inherent scale or meaning. Effect size is a cleaner measure of magnitude and should not be confused with statistical significance. Unfortunately, most HCI researchers have not yet embraced the use of effect sizes even though it is now mandated in many other scientific venues (e.g., American Psychological Association, 2010, p. 34). However, exemplary papers do exist, especially those performing meta-analyses on topics such as self-disclosure in digital environments (Weisband & Kiesler, 1996) or examining the influence of human-like faces in embodied agents on interaction experience (Yee, Bailenson, & Rickertsen, 2007), as well as individual experimental studies that compare effect sizes across conditions (Gergle et al., 2013).

Another way HCI researchers can better express magnitude is to report *confidence intervals* (Cumming & Finch, 2001; Smithson, 2003). Confidence intervals provide a more intuitive and meaningful description of the mean difference between the groups. Instead of providing a single number, they identify the range in which the true difference is likely to fall. Confidence intervals, and their corresponding confidence limits, are an intuitive way of specifying not just an estimate of the difference but also the likely minimum and maximum values of the difference. A good example drawn from a field experiment can be seen in Oulasvirta and colleagues' research (Oulasvirta, Tamminen, Roto, & Kuorelahti, 2005).

Finally, there is a more practical side to magnitude that is determined by the choice of experimental design and manipulations. Consider a study that shows a

[22] We use a two-condition example for ease of exposition.

large effect with a rather subtle manipulation vs. one that shows a large effect with an extreme manipulation. For example, demonstrating an increase in contributions to an online peer-production system by providing a graphical badge on a person's profile page (subtle) vs. paying them $100 to contribute more content (not-so-subtle). To the extent that you can produce the same size effects with the former, your results have greater magnitude, and oftentimes, practical importance.

Articulation. Articulation refers to the degree of detail that is reported about the research findings. Consider the following three descriptions which range from least to most detailed in discussing the results of a 3 (Input Style)×2 (Gender) factorial experiment: (a) "there was a significant performance difference between the three UI input styles"; (b) "There was a significant performance difference between input styles and also a significant performance difference by gender"; or (c) "There were significant differences between all types of input styles with style 1 being 75 % faster than style 2, which in turn was 18 % faster than style 3. Moreover, these performance differences across input styles were even stronger for females than for males, and females overall were 7.2 % faster than males." While the various statements are reporting the same general trend in findings, the last statement does so with much greater articulation. For a discussion of ways to enhance reporting of results with respect to articulation see Abelson (1995, pp. 104–131).

Generality. Generality represents the extent to which the research results apply outside the context of the specific study. One aspect of this is external validity, or the degree to which the results can be generalized to other situations, people, or times.

The sample and the population from which it is drawn often limits generality. For example, if you are only studying Facebook users, you cannot make claims that generalize to the entire world's population—especially given that a significant majority of the world does not actually use Facebook in a significant way. You can, however, make claims about the smaller population of Facebook users. Similarly, US college students, often easily recruited because they are required to serve in experiments as part of a course requirement, are not indicative of people in the rest of the world in many, many, ways.

Another limitation often comes from the choice of experimental and statistical controls employed in a study. In HCI, it is often the case that a highly controlled laboratory study with participants who have no history together may not be generalizable to the real-world field environment where the environment can be noisy and chaotic, people have prior relational histories, motivation can widely vary, etc. Using a wider range of contextual variations within studies, and a systematic program of replication and extension along with the application of meta-analysis (for an introduction to the technique, see Borenstein, Hedges, & Higgins, 2009; for HCI examples, see McLeod, 1992; Weisband & Kiesler, 1996; Yee et al., 2007) across numerous studies, are ways to broaden the scope of your findings and improve the generality of your research.

Interestingness. While the first three criteria can be treated in a more objective fashion, the last two have more subjective elements. Interestingness has to do with

the importance of the research findings, and this can be achieved in various ways. Here we focus on three dimensions of interestingness: theoretical, practical, and novelty.[23]

The *theoretical* dimension centers on experimental HCI research that seeks to inform. Theoretical contributions often consist of new or refined concepts, principles, models, or laws. For experimental work to be interesting on a theoretical dimension, the findings have to change what theorists think. If we consider theory as our best encapsulation of why things work as they do, then challenging that assumption or refining it in order to make our theories more complete or correct is a hallmark of good theoretical research. The extent to which the theory must change, or the number of theories that are influenced by your findings, are two key ways in which importance is assessed.

There are numerous experimental and quasi-experimental studies that make contributions on the theoretical dimension. For example, work by Zhu and colleagues challenges the traditional notion of online leadership, and suggests that it may be a more egalitarian construct than previously assumed (Zhu, Kraut, & Kittur, 2012; see also Keegan & Gergle, 2010). Dabbish and colleagues (Dabbish, Kraut, & Patton, 2012) used an innovative online experiment to reveal the communication behaviors and theoretical mechanisms by which commitment to online groups occurs. Finally, several classic studies in the domain of Fitts' Law have advanced the theory by demonstrating trajectory-based steering laws (Accot & Zhai, 1997; Wobbrock et al., 2008).

The *practical* dimension centers on experimental HCI research that seeks to solve everyday problems and issues. Practical contributions can take the form of the development of useful new metaphors, design guidelines or design patterns, new products or services, and design checklists or best practices. This type of work may take a more pragmatic and sometimes atheoretical approach to design and development. In these cases, experimental research techniques often focus on evaluating or verifying the utility of a new design or practice. Some excellent examples of this approach are provided in Kohavi and colleagues' work on using web experiments to inform design choices (Kohavi, Henne, & Sommerfield, 2007; Kohavi & Longbotham, 2007; Kohavi, Longbotham, & Walker, 2010).

The *novelty* dimension centers on experimental HCI research that seeks to invent. This often includes the design, development, and deployment of new systems; new infrastructures and architectures; and new tools or interaction techniques. While not all novel contributions of this type in the HCI literature require experimental support, many are accompanied by an experimental demonstration of their utility and how well they perform in new settings or relative to existing best practices or state-of-the-art algorithms or systems.

[23] While we separate these three areas in order to discuss the relative contributions that are made in each, it is not to suggest that these are mutually exclusive categories. In fact, some of the most influential work has all three dimensions. For a more nuanced discussion of the integration of theoretical (basic) and practical (applied) research in an innovation context see Stokes (1997) *Pasteur's Quadrant*.

Gutwin and Penner's work on telepointer traces (Gutwin & Penner, 2002), Wigdor and colleagues' LucidTouch system (Wigdor, Forlines, Baudisch, Barnwell, & Shen, 2007), or Zhai and Kristensson's work on the SHARK shorthand gesturing system (Kristensson & Zhai, 2004; Zhai & Kristensson, 2003) all make use of elements of experimental design[24] to rigorously demonstrate the utility of their novel designs and systems.

Credibility. Credibility is established by convincing the readers and reviewers that your work has been performed competently and with regard to common pitfalls and traps—it serves to bolster the plausibility of your claims. Much of what we have discussed throughout this chapter is aimed at establishing and supporting the credibility of your work. Doing things correctly, according to preestablished best practices and guidelines, is the easiest way to convince others of the credibility of experimental research. Dealing with internal and external validity, choosing a sample and understanding its limits, recognizing potential confounds, reporting on large and meaningful effects, performing appropriate analyses and correctly reporting and representing your findings are all keys to establishing credible experimental research.

Writing Up Experimental Research

In order for experimental HCI research to have an impact, it needs to be communicated to other researchers. While a detailed discussion of writing and dissemination is beyond the scope of this chapter—and several excellent guides already exist (e.g., Bem, 2003)—the following provides a brief description of the central elements required when reporting experimental research.

The general form of an experimental research article follows the hour-glass writing form. It is broad at the beginning and end, and narrow in the middle. Keep in mind that the main goal of your research paper is to motivate and detail your argument, demonstrate what you did, and convince the reader of your contribution. It is not a chronology of everything you did from day one, nor is it a detailed description of every single fact you uncovered. It is a pointed argument. The following presents a standard structure for an experimental research piece, and we focus on elements that we feel are often misreported or problematic in HCI related venues:

Introduction. The introduction should answer the question, "What is the problem?" and "Why should anyone care?"[25] It should provide an overview of the work and

[24] Not all of these studies are strict randomized experiments. For example, the SHARK evaluation does not make use of a control or comparison group. However, many use experimental research techniques to effectively demonstrate the feasibility of their approach.

[25] The framing questions in this section are drawn from Judy Olson's "10 questions that every graduate student should be able to answer." The list of questions and related commentary can be found here: http://beki70.wordpress.com/2010/09/30/judy-olsons-10-questions-and-some-commentary/

develop the central argument for the paper. It should identify the problem, provide rationale for why it matters and requires further research, describe and situate the research in the context of related literature, and end with the specific goals of the study often stated in the form of hypotheses or research questions. Be sure to state the research questions early, and walk the reader through your argument. Use plain English. Provide examples. Be concrete.

Method. The method section should aim to answer the question, "What did I do?" It should begin with a detailed description of who the *participants* were (e.g., age, gender, SES, education level, and other relevant demographic variables). It is also important to know about the motivations used to achieve participant involvement. Was it done for course credit? Were the participants paid? If so, did it depend on their performance? etc.

The *sampling procedure* should then be discussed. For example, were the participants drawn from a randomized national sample or perhaps snowball sampling was used? Next, the approach used to *assign participants* to experimental conditions should be described. Were the participants randomly assigned, was some form of paired assignment used, or were preexisting groups used (e.g., classrooms)?

The next area to include in the method is a description of the *experimental design* and the *experimental conditions*. The type of design should be clearly articulated (e.g., between- or within-subjects, mixed factorial design, or interrupted time series). The dependent and independent variables should also be described. This should be followed by a description of the *stimuli* and *materials* used to collect the data.

Finally, the written *procedure* should provide a detailed description of the processes used to collect the data. Describe any particular machinery, software, or measurement instruments. Discuss how the participants were handled before, during, and after the study, and detail the presentation order of materials. This should be followed by a description of the analysis where you detail what statistical comparisons were planned, discuss how missing data were treated, state how your dependent variable was captured, scored, annotated, etc.

The rule of thumb for the amount of detail that should go into the method section is that it should be enough for another researcher to be able to replicate the study if they chose to do so.

Results. The results section should aim to answer the question, "What did I find?" It should present the analyses performed and the major findings. You should present the results in a way that best supports the central argument being proposed in the paper, and be explicit when addressing central research questions and hypotheses.

The results section should focus on the most important findings or DVs. Remember, you are presenting results that are relevant to the central argument of the paper (both those that support and contradict your argument). Be sure to state each finding in a clear form without the use of jargon, and then support it with statistics. Remember that the statistics are not the focal point of the results section. The statement of the finding is the important part, and the statistics should be used to bolster the reader's confidence in that statement. Show the most relevant findings in tables

and figures, and be sure to point out the figures and tables in the accompanying prose. It can also be useful to interpret as you present, although you need to be clear about what are actual results and what are interpretations of the results. Finally, end the results section with a reminder of the purpose of the experiment and provide a light summary of the results with respect to the central argument.

Discussion. The discussion section should aim to answer the question, "What does all of this mean?" and "Why does it matter?" Remember, this section (and the paper as a whole) should be a pointed argument. The discussion section is where you can contextualize your results both with respect to the central research questions and hypotheses and in relation to prior work in the area.

In this section you should start by reviewing the evidence you have garnered toward your position and discuss the evidence against it. Be sure not to oversell your findings. You should also be sure to discuss the limitations of the current study or approach and address possible alternative explanations for your findings.

Once you have discussed your results in detail, you can begin to talk about the broader implications of the work whether they are for design, policy, or future work. You can describe the ways in which new experiments can be performed to address open questions or describe new directions that need to be addressed given the findings you have revealed.

Conclusion. Finally, you should conclude the paper with a restatement of your work. The conclusion is, in many ways, like the introduction of the paper. This is often a single paragraph that reminds the reader of the initial goals of the work, what you found, what it means, and why it matters—both for the particular questions under investigation as well as more broadly.

Personal Story about How the Authors Got into this Method

In this section we describe our personal experiences with research we conducted together with colleagues Randy Pausch and Peter Scupelli exploring the cognitive effects of physically large displays.

At the time we began our large display work (around 1999), LCD manufacturing was becoming significantly more efficient, creating a supply of cheaper and larger displays. Furthermore, projectors and digital whiteboards were becoming commonplace in conference rooms, and researchers were exploring the extension of these large displays into more traditional office spaces. Although researchers had articulated the qualitative benefits that large displays had on group work, little research had been done to quantify the benefits for individual users, which we had anecdotally noticed in our various new display setups. We thus set out to compare and understand the effects that physical display size (i.e., traditional desktop displays vs. large wall displays) had on task performance (Tan, Gergle, Scupelli, & Pausch, 2006).