Chapter 17
Experimental Design

This chapter deals with a general review of experimental design with emphasis on lab-based experiments. Generally speaking, this excludes nonexperimental designs (e.g., observational studies) and fieldwork. The point of this chapter is to review experimental procedures and methods, providing a framework, or context if you will, for eye tracking experiments. The content of this chapter reviews basic experimental designs, and does not differ greatly from texts used in introductory experimental psychology classes. For example, see Coolican (1996) for such a text. From this set of designs, this chapter identifies factorial designs as particularly popular in eye tracking research.

17.1 Formulating a Hypothesis

When designing an experiment, one of the first considerations should be the formulation of the research question. Formulating this question properly should lead the researcher toward a good design. For example, starting out with a statement such as, "I wonder what would happen if...," is what could be considered a naïve approach because it is not necessarily based on any assumptions or theories and does not identify any particular direction for testing. On the other hand, stating, "I bet this result would happen if...," already suggests an underlying assumption as well as potential candidate measures, e.g., some quantity that can be measured during experimental outcomes. The point is that a hypothesis is required when designing a formal experiment. Given a hypothesis, the experiment almost "designs itself" because it is then mainly concerned with accepting or rejecting the preliminary hypothesis, if it is stated with sufficient precision.

A.T. Duchowski, *Eye Tracking Methodology*, DOI 10.1007/978-3-319-57883-5_17

More formally, an experimental design is often drawn from the formulation of a *null hypothesis* (H_0), i.e., a statement predicting no difference in measured results collected between two (or more) sets of data obtained under different conditions. Hence, no effect is expected. The point of the experiment then is to reject the null hypothesis, showing that results are highly unlikely if the null hypothesis is true, thereby providing support for the alternative hypothesis. A classic example that is familiar to most people is that of a new drug being tested. The null hypothesis states that the drug has no effect, or more specifically, its effect is no different from a placebo (a sugar pill that is known not to have any effect). Establishing the hypothesis immediately suggests a logical course of action: how to administer the drug, and what to measure. The drug, or treatment, could be administered to one group of participants, with another group of participants receiving the placebo. Measurements can then be compared between the two groups. Using statistical analysis, if the measurements are no different, then the null hypothesis is accepted (indicating the new drug's inefficacy), otherwise, the null hypothesis is rejected. The latter case lends support to the statement that the drug has an effect. Note, however, that this support does not constitute absolute proof. The experiment, rooted in the conventional scientific method, provides scientific evidence for the drug's effect, but not proof of its effectiveness. This perhaps subtle distinction between scientific evidence and proof is too often ignored by students and overzealous marketing agencies.

Eye tracking studies generally do not involve pills or other digestibles. Instead, the "treatment" is often some differing form of interactive display, e.g., the computer's graphical user interface, or GUI (pronounced "gooey"), or varying forms of visual stimulus, e.g., two different images. In most cases, study participants are often given fairly specific tasks such as execution of some function, e.g., open a Web browser or find a specific GUI icon. Measurements then include reaction times (how quickly participants perform the action, on average), error rates (how many mistakes occurred, on average), and, of course, measures related to participants' eye movements. The latter usually include fixations, fixation durations, etc. How the different conditions are manipulated within the experiment is governed by the study's experimental design.

More formally, the treatment being manipulated or changed in value is referred to as the Independent Variable, or IV. All other variables are held constant (or attempted to be held constant; variables outside the experiment's control affecting the measured outcome may confound the outcome and are known as confounding variables). Whatever is being measured (e.g., reaction time) is usually whatever is expected to be affected by the IV, and is known as the Dependent Variable, or DV. That is, the DV depends on the manipulation of the IV.

The remainder of this chapter reviews different types of experimental inquires that can be made to test a given hypothesis. Given the choice of one design over another, the chapter then provides a review of basic statistical tools that can be used to measure differences and hence the effect of the conditions under examination.

17.2 Forms of Inquiry

Coolican (1996) defines *investigation* as a general term for any study that seeks information (usually to test a hypothesis). An *experiment* is a particular form of study where, in general, all possible causes of variation in the effect being measured are eliminated except the one influence under investigation. The general rule of thumb is to vary one thing while keeping everything else constant. Ensuring that all other conditions are equal except the main effect suggests gaining control of the experiment. This is the key concern of experimental designs: how to ensure that only one condition is varied and all else is held constant. This may sound simple but it is not. Even in fairly highly controlled settings such as laboratories, there are still many factors that may influence the outcome of experiments. Simple and mundane considerations such as whether study participants performed their given tasks before or after lunch may matter. The degree to which conditions are controllable will determine the type of experiment (or nonexperiment) being conducted.

There are a few different dimensions that specify different forms of experimental designs, including:

- Experiments versus observational studies
- Laboratory versus field research
- Idiographic versus nomothetic research
- Sample population versus single-case experiment versus the case study
- Within-subject (repeated measures) versus between-subjects designs.

17.2.1 Experiments Versus Observational Studies

The distinction between experiments and nonexperimental observational studies revolves about the manipulation of an independent variable. Observational studies are generally made by observation without manipulation of an IV (e.g., consider gender as an IV; it cannot be manipulated). Being able to manipulate an IV is generally a prerequisite for the design of an experiment. Furthermore, in the interest of replicability, experiments often follow a standardized procedure. Variables, independent and dependent, need to be strictly defined, procedures undertaken during experimental trials need to be detailed, and results from analysis must be effectively reported. Most research papers follow a fairly similar format, partially so that other researchers can reproduce their experiments and (it is hoped) replicate their results. This format often includes:

1. Hypothesis: the null or alternative hypothesis, with theoretical justification for any given assumptions.
2. Design: which experimental design is ultimately chosen, is it a nonexperimental observational study, or if an experiment, what are the IVs and DVs, and how are participants grouped, if at all (e.g., within-subjects or between-subjects; see below).

3. Participants: the number of participants in the study, with demographic data such as age ranges and gender distribution (all reported anonymously).
4. Apparatus: the devices used; in eye tracking studies, one generally reports the operating characteristics of the eye tracker including its underlying mechanism (e.g., video-based, combined pupil–corneal reflection), accuracy (e.g., 0.5°), sampling rate (e.g., 50 Hz), operating range (e.g., 50 cm), and whether any other auxiliary devices such as chin rests are needed.
5. Procedures: essentially what is told to participants prior to and following their experimental trials; is there any training or instructions (usually read from a script), what type of calibration is used, etc.
6. Tasks: what do the participants actually do? Task definition is particularly important, more so for eye tracking studies because eye movements are known to be task-dependent (gaze is simultaneously bottom-up, stimulus-driven as well as top-down, goal-oriented).

17.2.2 Laboratory Versus Field Research

Conducting an experiment in the laboratory can often allow greater control over experimental conditions than what can normally be achieved in the field. Control is probably the chief reason for holding experiments in the laboratory. Indeed, for various computer-related experiments, such as usability testing, numerous usability labs have appeared with specialized recording "studios" equipped with one-way mirrors, video cameras, and eye trackers. Detractors of lab experiments question the generalizability of results to less artificial settings of one's office, home, etc. In a nutshell, laboratory experiments suffer from a reduction of ecological validity but, through increased control, gain internal validity.

For eye tracking research, equipment often dictates pragmatic constraints such as whether the experiment needs to remain in the lab or whether the eye tracker can be used out "in the field". With increasingly smaller and more portable equipment, eye tracking experiments need not be confined to the lab. For example, table-mounted eye tracking equipment can be fairly easily transported and with a laptop experiments can be conducted "on-site". Head-mounted gear is also becoming increasingly less cumbersome and more affordable (Li et al. 2006) and hence can be used for various experiments performed outside the lab.

17.2.3 Idiographic Versus Nomothetic Research

This distinction pertains to the study of an individual (idiographic) versus the study of larger populations. Generally speaking, beyond clinical evaluations of individuals, or evaluation of custom-built solutions, eye tracking studies seek to uncover similarities of viewing patterns of large groups of viewers (e.g., over art, or

computer-generated scenes), even though variability and task-dependence of eye movements are widely acknowledged. In a nomothetic approach, important concerns are generalizability of results to larger populations and the selection of appropriate population representatives.

A particularly instructive example of a (difficult) field study involving a specific (somewhat idiographic in spirit) population was presented by Hornof and Cavender (2005) who investigated an interactive eye drawing application designed for children with severe motor impairments. In a pilot study, Hornof and Cavender (2005) first employed ten participants without disabilities where half were children (average age of 12; the other half's average age was 26). (Prior to this, the authors performed two user observation studies, one with children and adults without disabilities, the other with adults with severe motor disabilities.) Some users were based locally (presumably in the lab), others were at remote locations. The final evaluation study was performed by four participants from the target audience, aged 9, 12, 18, and 61. The difficulty here is of course selection of a representative sample population. Initially, it makes sense to perform prototypical evaluations in a highly controlled environment (lab) with population samples that may not necessarily generalize to the target population (e.g., adults without disabilities). Although both constraints present generalizability problems in terms of environment and individuals' abilities, this operational constraint makes sense: gross problems can be identified early in the prototypical development stages, before moving on to field trials with members of the target audience.

17.2.4 Sample Population Versus Single-Case Experiment Versus Case Study

Selection of a sample population is generally performed with the intention of generalizing results to a wider (if not potentially global) population. There are instances when it is appropriate to consider a population consisting of a single individual. This may be a case study of an individual, as is often reported in clinical accounts. Phineas P. Gage presented the famous case of an individual accidentally lobotomized by an iron rod propelled through his frontal brain regions in a rock blasting accident. What was interesting about this case was the change in Gage's personality following the accident. Prior to this accident Gage was characterized as mild-mannered, however, following the accident he had become aggressive, rude, and "...indulging at times in the grossest profanity (which was not previously his custom), manifesting but little deference for his fellows, impatient of restraint or advice when it conflicts with his desires...," according to the physician Harlow in 1868 (Harlow 1868). Analogously, there may be highly interesting and informative eye tracking clinical cases.

An alternative to the case study is the single-case experiment. This is a quasi-experimental design in that it lacks randomized allocation of participants to treatment conditions. This design is appropriate for specific individuals or small groups of

people, e.g., experts, as they may be employed in heuristic usability evaluations. Performance of the expert(s) can then be compared to the average performance of novices as a type of baseline comparison under similar or identical conditions. This type of approach may be taken in training and assessment studies.

17.2.5 Within-Subjects Versus Between-Subjects

Of the many experimental approaches available, the two most likely methods of collecting data from groups is either via a within-subjects (repeated measures) or via a between-subjects design. A within-subjects design uses one group of participants and tests them under all treatment conditions. A between-subjects design uses different groups of participants, where different treatments are assigned to different groups.

The within-subjects design repeats treatments per individual, hence it is also known as the repeated-measures design. The most prominent problem with repeated measures is that the analysis of results will suffer from order effects unless care is taken to counterbalance the conditions. Order effects may include fatigue or learning, for example. To counterbalance the conditions, a Latin square may be used to (randomly) assign conditions to participants. The Latin square of order n is an $n \times n$ array in which each cell of the array contains one of a set of n symbols such that each symbol occurs only once in each row and column; e.g., given $n = 4$ conditions, A, B, C, D, the following Latin square is generated,

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

where each row can be used to assign a treatment sequence to each of four study participants.

A between-subjects design can avoid the repetition of treatments by using different groups of participants per treatment. For a four-condition experiment, four groups could be used, with different people assigned to each group. Care has to be taken to avoid accidental homogeneity of groups, e.g., testing two groups where one group is entirely male, the other entirely female, introduces gender bias into the results. Random assignment may not attenuate participant variability fully, however. Other strategies for group assignment involve prescreening of participants, and/or targeted assignment by representation. The former involves some form of participant assessment, e.g., questionnaire or pretest. The latter may involve assignment ensuring roughly equal representation of disparate individuals in each group (e.g., equal representation of people with 20/20 vision and myopes in each group).

There are two disadvantages to the between-subjects design. First, more subjects are needed to obtain similar power to a within-subjects design. Recruiting and running more subjects can be costly and time consuming. Second, if there is too much variance among the participant groups, statistical analysis may be complicated; e.g., it may not be possible to perform parametric evaluation of the statistical mean differences.

17.2.6 Example Designs

A general discussion of experimental designs is complicated without a specific context and specification of IVs/DVs being tested. It is often easier to settle on the IVs/DVs first and then develop the design. Doing so will stipulate the requirements for the number of groups or experimental trials required, depending generally on whether a within- or between-subjects design is adopted. In this section some basic designs are offered. Not all are applicable to eye tracking studies, but they are given for a better sense of completeness.

Single Individual, Time Series

This design is exemplified by the traditional drug effectiveness experiment. In this design, there is just one participant, and measurements are taken over time. Prior to administration of a treatment, a baseline measurement is taken. Subsequent measurements are then compared to the baseline to test for the drug's effect. Figure 17.1 is an example of a simple *ABAB* type design where *A* indicates no treatment (baseline) and *B* indicates treatment. In Fig. 17.1 only *ABA* is shown. The dashed lines indicate administration and subsequent cessation of treatment. The graph itself would suggest some level of the dependent variable being measured (perhaps some form of subjective well-being or alternatively some physically measurable indicator such as blood pressure). Analysis of this type of experiment typically requires comparison of the DV mean during the specific time sequences during which it was known the drug was taken, e.g., time period over which treatment *B* was active. If this mean is statistically different from the mean during the period when the drug is not present, i.e., period of treatment *A*, then the drug is said to have an effect.

Note that this design can be thought of as either one with two independent variables, *A* and *B*, or just one but with two levels. Relabeling the diagram in Fig. 17.1 with $A_\circ$ and A_1 would indicate one treatment, or *factor*, administered at two levels: $A_\circ$ meaning absent, A_1 meaning present. This would now be considered a single factor design.

General Pre–Post Design

The single individual design can be adapted to a single group design, wherein the same group of individuals is given the treatment in the *AB* treatment administration sequence, and the data are now analyzed over the entire sample of participants. This design is known as pre–post because measurements are taken prior to and following treatment administration. This design is shown in block form, along with other designs, in Fig. 17.2. Once again, this too can be considered a single factor design.

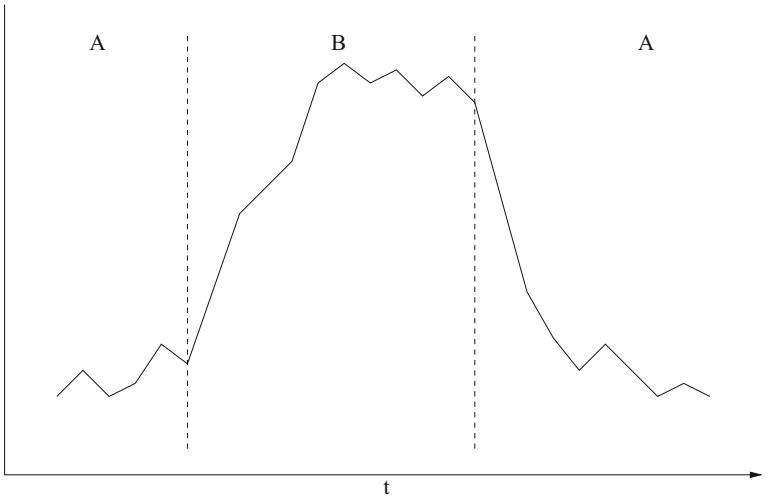

Fig. 17.1 Example of single-subject, time series *ABAB* type design

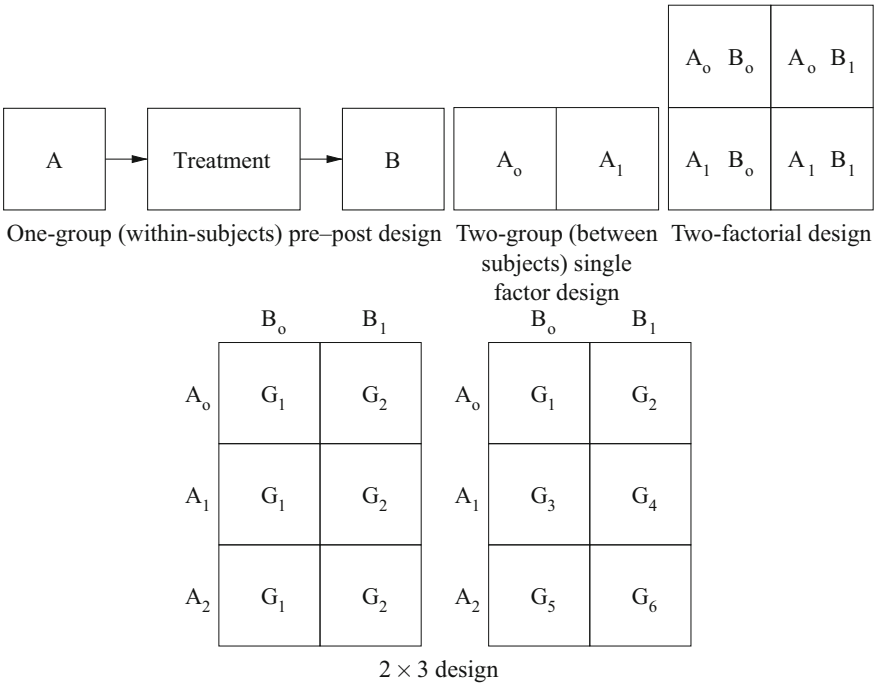

Fig. 17.2 Factorial design diagram examples

Two-Group Design

The previous single factor (one experimental variable) designs are within-subjects because all participants (including the individual, naturally) are given the treatment. Extending those designs to between-subjects leads to a two-group factorial design, shown in block form in Fig. 17.2. Each group, identified by the boxes, receives either $A_\circ$ or A_1, where $A_\circ$ is a placebo and A_1 is the treatment. Being a between-subjects design, care must be taken in assignment of participants to groups. This is accomplished via randomization, prescreening, or targeted assignment, as discussed above. Comparison of the mean response to the stimulus (A_1) versus the mean response to the placebo ($A_\circ$) would determine significance (or insignificance) of the treatment (i.e., its effect).

Two-Factorial

Consider the four boxed two-factorial design in Fig. 17.2. There are four experimental conditions, composed of a combination of two experimental variables, each administered at one of two levels. This is a 2×2 factorial design, where the 2 indicates the number of levels of each factor. What is important is how groups of participants are allocated to each condition. Several permutations are possible:

- If each box in the 2×2 cube is assigned a unique group of participants, this design is a 2×2 four-group, or fully independent, between-subjects (or just 2×2 between-subjects) design.
- If we assigned one group of participants to condition $B_\circ$ (the left column) and another group to condition B_1 (right column), then the design becomes a 2×2 mixed design (because it mixes within- and between-subjects). Because the same group would receive treatment $B_\circ$ or B_1, repeated measures analysis would be required to test for significance between these factorial levels inasmuch as the measurements of B are no longer independent.

2×3 *Factorial Design*

Finally, consider the 2×3 design shown at the bottom of Fig. 17.2 in yet another form of representation, where groups of participants are now made explicit with the subscripted G notation. Both are 2×3 designs because there are two levels of B and three levels of A. The design at left, however, differs because treatment A varies within-subjects. Analysis of A's effect, therefore, examined across rows, will require repeated measures, or nonindependent analysis across the rows of the data collected. Analysis of B's effect, examined across columns, however, can be performed via independent analysis methods because this treatment is administered between-subjects. In contrast, the 2×3 design at right is fully independent, or fully between-subjects.

17.3 Measurement and Analysis

Experimental design is generally concerned with the manipulation of independent variables and the subsequent measurement of dependent variables. The general expectation (hypothesis) is, of course, that manipulation of the IV will render some effect on the DV and that, most importantly, the effect will be measurable and (statistically) significant. The "trick" to conducting a meaningful experiment is the operational definition of the dependent variable, that is, that some meaningful measurement of the expected effect can be defined.

To give an example related to eye tracking, we could test the attentional quality of a banner advertisement on a Web page. To do so, we hypothesize that a dynamic blinking banner ad is more visually attractive than a static image. This hypothesis is based on the underlying theory of low-level vision which tells us that vision, particularly in the periphery, is sensitive to "sudden onset" stimuli. Thus we have the expected cause and effect that we operationalize by first defining the IV to be the static or dynamic nature of banner ad, i.e., presence of motion in the stimulus (one can be much more specific and stipulate such details as the frequency of the animation, its size, position on the Web page, and so on; this would lead to an increase in the number of IVs being manipulated).

The definition of the DV should match our initial qualitative description of "attentional quality". We can do so by quantifying visual attention in terms of the number of fixations devoted to the banner ad during a given interaction (Web surfing) session. Note that this operationalization of the DV is quite specific in its assumption of fixations denoting visual attention. A common criticism of this assumption is that it is possible to voluntarily disengage attention from foveal vision. This is true, but because we cannot measure this covert mechanism of attention, the best we can do is acknowledge awareness of this fact and stipulate that we are only measuring overt attention and assuming that during the course of our study we expect participants' attention to be aligned with foveal vision.

Other important details must also be addressed. In particular, the task that participants are going to perform needs to be defined. For example, navigation or search may affect the Web task and hence recorded eye movements differently. Additionally, remaining procedural details must also be specified; e.g., how long should individual trials last, how many trials should each individual perform, etc. (These will to a certain extent depend on the complexity of the design in terms of number of IVs.) Specific to eye tracking, calibration is important: how many points are to be used, and how often is calibration to be performed?

The test of a hypothesis depends on a comparison of the outcomes of the dependent variables, following manipulation of the IVs. In the banner ad example, the DV was operationalized quantitatively as a number of fixations. Given the two conditions, the static and dynamic ad, we would obtain two groups of measurements: number of fixations per participant group per condition. Because there are only two outcomes, and we have numerical data, we can obtain descriptive statistics about the two results and perform parametric tests on the outcome differences. For example, as is fairly

common, means and variances would be computed for both groups of fixations. Subsequently, a t-test could be made on the pair of means to test for statistical significance in the difference of means.

The t-test basically reports whether the two sets of numbers overlap in a statistical sense, based on the assumption of normality of the data. That is, the data are assumed to fit a normal (Gaussian) distribution. The t-test checks whether the two distributions overlap. If they do not, one can then claim, at a sufficient level of significance as reported by the t-test, that the outcomes differ, and therefore, an effect is observed. In the hypothetical banner ad example, if it turned out that the number of fixations falling on the dynamic ad was significantly greater than the number of fixations falling on the static ad, then we could state that our data support the initial hypothesis that the dynamic blinking ad is more visually attractive by drawing to it a significantly higher number of fixations. (More formally, the null hypothesis H_0 would have been given as one stating no expected difference between outcomes, which would have to be rejected in favor of our stated alternative hypothesis that the outcomes do differ.)

Given more than two experimental conditions, e.g., dynamic ad, static ad, no ad (a control condition), with all other experimental procedures remaining unchanged, the statistical parametric test that is used quite often as a test for significance is the Analysis Of Variance, or ANOVA. ANOVA is conceptually an extension of the t-test in that it only tests for overlap of the data distributions (assumed to be normal). Its main function is to report whether there is significant distance between overlap of the means. For example, given our three ad conditions, let's say we collected five fixations per group (and recorded them as either within the banner ad or not). We would then plug the resultant data ($n = 15$) from the three groups into a statistical program such as SPSS, S-Plus, or R and obtain an ANOVA table. To find out whether there is sufficient dispersion in the means, i.e., a significant difference between the means, the F statistic and level of significance are examined. In this example, suppose $F = 4.761$ at significance level 0.030. This result is significant, and would be reported as $F(2, 12) = 4.761, p < 0.05$, where the numbers in parentheses following F indicate the between- and within-group degrees of freedom (df), respectively, whose total should add up to $n - 1$. In this case there were three groups and $n = 15$, giving between-groups df $= 2$ and within-groups df $= 12$ so that $2 + 12 = n - 1$.

ANOVA is a particularly popular analysis tool and is used as a preliminary indicator of (statistically significant) effect. However, it only reports a difference among the means. To pin down which mean (and hence condition) differs from all others (perhaps all conditions differ significantly from each other), usually a pairwise comparison of means is then required, such as the pairwise t-test or the pairwise Kruskal–Wallis test if the data cannot be assumed to be normally distributed (or the measurement scale of the data is not an equal interval scale or the sampled data do not have approximately equal variances).

In general, the type of statistical test of difference performed depends on the type of data being collected and the number of samples (groups) measured. Table 17.1 lists statistical tests of difference for sample pairs. Table 17.2 lists statistical tests of difference for multivariate data, in this context meaning two or more variable quantities. Nominal data are usually not measured, and refer to sorting or assignment

Table 17.1 Statistical tests of difference of sample pairs (df = 1)

	Samples	
Measurement level	Independent	Nonindependent
Nominal	Chi-square	(Binomial) Sign test
Ordinal	Mann–Whitney U test	Wilcoxon signed ranks
Parametric	*z*-test, *t*-test for independent means	*t*-test

Table 17.2 Statistical tests of difference of multivariate data (df > 1)

	Samples	
Measurement level	Independent	Nonindependent
Nonparametric	Chi-square	Kruskal–Wallis test
Parametric	ANOVA	ANOVA

of values into categories, hence they are also known as categorical data (e.g., number of blue-eyed people in the sample population). Ordinal data are measured but only denote the order or position of a data point (e.g., the top five scorers on a midterm exam).

Generally speaking, eye movement data are considered parametric because related metrics can be represented by a uniform (equal distance) interval/ratio scale. An interval scale is one composed of equal units, where, for example, the distance from 160 to 165 cm is the same (*means* the same) as that between 170 and 175 cm. Related to the interval scale is the ratio scale, where the latter is an interval scale with a necessary and absolute zero. Examples of ratio scales include reaction times or distance. Timing starts from zero, and in this scale it makes sense to say that if something completes in half the time of something else, then it is twice as fast. Note that in the above eye movement example, the number of fixations can only be considered on this scale if it makes sense to say that twice as many fixations counted on one ad is in some sense twice as meaningful or valuable as on the other (and zero fixations is also meaningful). Fixations can be interpreted in this way if we consider them as indicators of cognitive load; e.g., devoting twice as many fixations to a region may mean that twice as much cognitive effort is being exerted (an operational assumption). This might not always be a valid interpretation, however. For example, twice as many fixations on some conspicuous portion of the screen (e.g., an ad) may mean the viewer is bored or distracted by something entirely different from the ad content, and hence is not indicative of cognitive load at all (just the opposite!). Thus one must be very careful in considering the type of measurement being recorded, and how the variable is being operationalized.

17.4 Summary and Further Reading

This chapter attempted to provide a rather general overview of experimental design. Being a very broad topic in itself, the resulting summary is most certainly incomplete. However, the main points that appear to be relevant to eye tracking are as follows.

1. Start any experiment with a good hypothesis statement; this is a good strategy for most experimental designs, not only ones concerned with eye tracking. For eye movement work specifically, think in terms of the kind of data that the tracker is able to provide, e.g., fixations, fixation durations, etc.
2. This summary tended to favor experiments over observational studies, however, eye tracking observational studies are not ruled out by any means. It is quite conceivable that observational studies can be performed to formulate an initial impression of how people may look at some visual stimulus. For example, eye movements collected over a prototype of an interface may guide the designer on the interface's layout. Similarly, layouts of Web pages, advertisements, or other visually rich media may be evaluated in this manner.
3. Due to the nature of the equipment, most eye tracking studies are lab-based, although due to the emergence of cheap head- or body-worn equipment, one can conduct eye tracking studies in the field. For a fascinating example of field studies (literally!), see the eye tracking work with lemurs by Shepherd et al. (2004). (It seems lemurs tend to look at the tails and heads of other lemurs.)
4. The issue of idiographic versus nomothetic study, similar to sample population versus single-case versus case study design depends on the situation (as most experimental design considerations do anyway). The point is, it would certainly be desirable to generalize eye movement behavior to large populations, e.g., all Web users. However, in some cases, this may not make sense. As in Cavender's (2005) studies, the target population was quite specific and therefore a restricted sample population made sense. A good guideline here would be to consider who the target audience is in selecting the sample population during the design of the experiment.
5. The experimental design review tends to favor factorial designs. Indeed, a good deal of eye tracking research appears to employ these designs. Whether they are conducted within- or between-subjects depends on the operationalization of the independent and dependent variables and other constraints of the experiment (e.g., time and money being the notorious ones).
6. The review of data analysis techniques attempted to be as complete as the competing goal of being succinct would allow. The general (and potentially dangerous) recommendation is the use of ANOVA as the main test of multivariate statistical difference. If ANOVA reports statistical significance, ensuing pairwise t-tests or Kruskal–Wallis tests should follow. The potential danger with this advice is that ANOVA may then be seen as a hammer with all experimental data nails. Although ANOVA is popular in the eye movement literature, the most conservative advice that can be given is to use the right tool for the job: carefully select the appropriate statistical analysis tool for the data collected.

The reader is strongly encouraged to delve into the rich literature on experimental design and/or attend experimental psychology lectures, if at all possible (inspiration for this chapter grew from personal notes taken during two guest lectures given by psychologist Eugene H. Galluscio in my eye tracking methodology course at Clemson University, Fall 1999). A good place to start is Coolican's (1996) introductory text.

Beyond experimental psychology references, especially for human–computer interaction work, where eye tracking is by now fairly commonplace, consider (Stone et al.'s 2005) book on user interface (UI) evaluation or (Dix et al.'s 2004) text on human–computer interaction (HCI).