3 From Vague Idea to Experimental Design

In Chapter 2, we described the competencies needed to build, evaluate, use and manage eye-trackers, as well as the properties of different eye-tracking systems and the data exiting them. In Chapter 3 we now focus on how to initially set up an eye-tracking study that can answer a specific research question. This initial and important part of a study is generally known as 'designing the experiment'.

Many of the recommendations in this chapter are based on two major assumptions. First, that it is better to strive towards making *the nature of the study experimental*. Experimental means studying the effect of an *independent variable* (that which, as researchers, we directly manipulate—text type for instance) on a *dependent variable* (an outcome we can directly measure—fixation durations or saccadic amplitude for instance) under tightly controlled conditions. One or more such variables can be under the control of the researcher and the goal of an experiment is to see how systematic changes in the independent variable(s) affect the dependent variable(s). The second assumption is that many eye-tracking measures—or dependent variables—*can be used as indirect measures of cognitive processes that cannot be directly accessed*. We will discuss possible pitfalls in interpreting results from eye-tracking research with regard to such cognitive processes. Throughout this chapter, we will use the example of the influence of background music on reading (p. 5). We limit ourselves to issues that are specific to eye-tracking studies. For more general textbooks on experimental design, we recommend Gravetter and Forzano (2008); McBurney and White (2007), and Jackson (2008).

This chapter is divided into five sections.

- In Section 3.1 (p. 66) we outline different considerations you should be aware of depending on the rationale behind your experiment and its purpose. There is without doubt huge variation in the initial starting point depending on the reason for doing the study (scientific journal paper or commercial report, for instance). Moreover, the previous experience of the researcher will also determine where to begin. In this section we describe different strategies that may be chosen during this preliminary stage of the study.
- In Section 3.2, we discuss how the investigation of an originally vague idea can be developed into an experiment. A clear understanding is needed of the total situation in which data will be recorded; you need to be aware of the potential causal relationships between your variables, and any extraneous factors which could impact upon this. In the subsections which follow we discuss the experimental task which the participants complete (p. 77), the experimental stimuli (p. 79), the structure of the trials of which the experiment is comprised (p. 81), the distinction between within-subject and between-subject factors (p. 83), and the number of trials and participants you need to include in your experiment (p. 85).
- Section 3.3 (p. 87) expands on the statistical considerations needed in experimental research with eye tracking. The design of an experiment is for a large part determined by the statistical analysis, and thus the statistical analysis needs to be taken into consideration during the planning stages of the experiment. In this section we describe

how statistical analysis may proceed and which factors determine which statistical test should be used. We conclude the section with an overview of some frequently used statistical tests including for each test an example of a study for which the test was used.

- Section 3.4 (p. 95) discusses what is known as *method triangulation*, in particular how auxiliary data can help disambiguate eye-tracking data and thereby tell us more about the participants' cognitive processes. Here, we will explore how other methodologies can contribute with unique information and how well they complement eye tracking. Using *verbal data* to disambiguate eye-movement data is the most well-used, yet controversial, form of methodological triangulation with eye-movement data. Section 3.4.8 (p. 99) reviews the different forms of verbal data, their properties, and highlights the importance of a strict method for acquiring verbal data.

3.1 The initial stage—explorative pilots, fishing trips, operationalizations, and highway research

At the very outset, before your study is formulated as a hypothesis, you will most likely have a loosely formulated question, such as "How does listening to music or noise affect the reading ability of students trying to study?". Unfortunately, this question is not directly answerable without making further operationalizations. The operationalization of a research idea is the process of making the idea so precise that data can be recorded, and valid, meaningful values calculated and evaluated. In the music study, you need to select different levels or types of background noise (e.g. music, conversation), and you need to choose how to measure reading ability (e.g. using a test, a questionnaire, or by looking at reading speed). In the following subsections, we give a number of suggestions for how to proceed at this stage of the study. The suggested options below are not necessarily exclusive, so you may find yourself trying out more than one strategy before settling on a particular final form of the experiment.

3.1.1 The explorative pilot

One way to start is by doing a *small-scale explorative pilot study*. This is the thing to do if you do not feel confident about the differences you may expect, or the factors to include in the real experiment. The aim is to get a general feeling for the task and to enable you to generate plausible operationalized hypotheses. In our example case of eye movements and reading, take one or two texts, and have your friends read them while listening to music, noise, and silence, respectively. Record their eye movements while they do this. Then, interview them about the process: how did they feel about the task—how did they experience reading the texts under these conditions? Explore the results by looking at data, for instance, look at heat maps (Chapter 7), and scanpaths (Chapter 8). Are there differences in the data for those who listened to music/noise compared to those who did not? Why could that be? Are there other measures you should use to complement the eye-tracking data (retention, working memory span, personality tests, number of books they read as children etc.). It is not essential to do statistical tests during this pilot phase, since the goal of the pilot study is to generate testable hypotheses, and not a p-value (nevertheless you should keep in mind what statistics would be appropriate, and to this end it might be useful to look for statistical trends in the data). Do not forget that the hypotheses you decide upon should be relevant to theory—they should have some background and basis from which you generate your predictions. In our case of music and eye movements whilst reading, the appropriate literature revolves around reading research and environmental psychology.

3.1.2 The fishing trip

You may decide boldly to run a larger pilot study with many participants and stimuli, even though you do not really know what eye-tracking measures to use in your analyses. After all, you may argue, there are many eye-tracking measures (fixation duration, dwell times, transitions, fixation densities, etc.), and some of them will probably give you a result. This approach is sometimes called *the fishing trip*, because it resembles throwing out a wide net in the water and hoping that there will be fish (significant results) somewhere. A major danger of the fishing trip approach is this: if you are running significance tests on many eye-tracking measures, a number of measures will be significant just by chance, even on completely random data. If you then choose to present such a selection of significant effects, you have merely shown that at this particular time and spot there happened to be some fish in the water, but another researcher who tries to replicate your findings is less likely to find the same results. More is explained about this problem on p. 94.

While fishing trips cannot provide any definite conclusions, they can be an alternative to a small-scale explorative study. In fact, the benefits of this approach are several. For example, real effects are replicable, and therefore you can proceed to test an initial post-hoc explanation from your fishing trip more critically in a real experiment. After the fishing trip, you have found some measures that are statistically significant, have seen the size of the effects, and you have an indication of how many participants and items are needed in the real study. There are also, however, several drawbacks. Doing a fishing-trip study involves a considerable amount of work in generating many stimulus items, recruiting many participants, computing all the measures, and doing a statistical analysis on each and every one (and for this effort you can not be *certain* that you will find anything interesting).

It should be emphasized that it is *not* valid to selectively pick significant results from such a study and present them as if you had performed a focused study using only those particular measures. The reason is, you are misleading readers of your research into thinking that your initial theoretical predictions were so accurate that you managed to find a significant effect directly, while in fact you tested many measures, and then formulated a post-hoc explanation for those that were significant. There is a substantial risk that these effects are spurious.

3.1.3 Theory-driven operationalizations

Ideally, you start from previous theories and results and then form corollary predictions. This is generally true because you usually start with some knowledge grounded in previous research. However, it is often the case that these predictions are too general, or not formulated as testable concepts. Theories are usually well specified within the scope of interest of previous authors, but when you want to challenge them from a more unexpected angle, you will probably find several key points unanswered. The predictions that follow from a theory can be specified further by either referring to a complementary theory, or by making some plausible assumptions in the spirit of the theory that are likely to be accepted by the original authors, and which still enable you to test the theory empirically.

If you are really lucky, you may find a theory, model, statement, or even an interesting folk-psychological notion that directly predicts something in terms of eye-tracking measures, such as "you re-read already read sentences to a larger extent when you are listening to music you like". In that case, the conceptual work is largely done for you, and you may continue with addressing the experimental parameters. If the theory is already established, it will also be easier to publish results based on this theory, assuming you have a sound experimental design.

3.1.4 Operationalization through traditions and paradigms

One approach, similar to theory-driven operationalizations, is the case where the researcher incrementally adapts and expands on previous research. Typically, you would start with a published paper and minimally modify the reported experiment for your own needs, in order to establish whether you are able to replicate the main findings and expand upon them. Subsequently you can add further manipulations which shed further light on the issue in hand. The benefits are that you build upon an accepted experimental set-up and measures that have been shown in the past to give significant results. This methodology is more likely to be accepted than presenting your own measures that have not been used in this setting before. Furthermore, using an already established experimental procedure will save you time in not having to run as many pilots, or plan and test different set-ups.

Certain topics become very influential and accumulate a lot of experimental results. After some time these areas become *research traditions* in their own right and have well-specified *paradigms* associated with them, along with particular techniques, effects, and measures. A paradigm is a tight operationalization of an experimental task, and aims to pinpoint cause and effect ruling out other extraneous factors. Once established, it is relatively easy to generate a number of studies by making subtle adjustments to a known paradigm, and focus on discovering and mapping out different effects. Because of its ease of use, this practice is sometimes called 'highway research'. Nevertheless, this approach has many merits, as long-term systematicity is often necessary to map out an important and complex research area. You simply need many repetitions and slight variations to get a grasp of the involved effects, how they interact, and their magnitudes. Also, working within an accepted research tradition, using a particular paradigm, makes it more likely that your research will be picked up, incorporated with other research in this field, and expanded upon. A possible drawback is that the researcher gets too accustomed to the short times between idea and result, and consequently new and innovative methods will be overlooked because researchers become reluctant of stepping outside a known paradigm.

It should be noted that it is possible to get the benefits of an established paradigm, but still address questions outside of it; this therefore differentiates paradigm-based research from theory-driven operationalizations. Measures, analysis methods, and statistical practices, may be well developed and mapped out within a certain paradigm designed for a specific research tradition, but nothing prohibits you from using these methods to tackle other research questions outside of this area. For example, psycholinguistic paradigms can be adapted for marketing research to test 'top-of-the-mind' associations (products that you first think of to fulfil a given consumer need).

In this book, we aim for a general level of understanding and will not delve deeper into concerns or measures that are very specific to a particular research tradition. The following are very condensed descriptions of a few major research traditions in eye tracking:

- *Visual search* is perhaps the largest research tradition and offers an easily adaptable and highly informative experimental procedure. The basic principles of visual search experiments were founded by Treisman and Gelade (1980) and rest on the idea that effortful scanning for a target amongst distractors will show a linear increase in reaction time the larger the set size, that is, the more distractors present. However, some types of target are said to 'pop out' irrespective of set size; you can observe this for instance if you are looking for something red surrounded by things that are blue. These asymmetries in visual search times reflect the difference between *serial* and *parallel* processing respectively—some items require focused attention and it takes time to bind their properties together, other items can be located pre-attentively. Many manipulations of the basic visual search paradigm have been conducted—indeed any experi-

ment where you have to find a pre-defined target presented in stimulus space is a form of visual search—and from this research tradition we have learned much about the tight coupling between attention and eye movements. Varying the properties of targets and distracters, their distribution in space, the size of the search array, the number of potential items that can be retained in memory etc. reveals much about how we are able to cope with the vast amount of visual information that our eyes receive every second and, nevertheless, direct our eyes efficiently depending on the current task in hand. In the real world this could be baggage screening at an airport, looking for your keys on a cluttered desk, or trying to find a friend in a crowd. Although classically visual search experiments are used to study attention independently of eye movements, visual search manipulations are also common in studies of eye guidance. For an overview of visual search see Wolfe (1998a, 1998b).

- *Reading research* focuses on language processes involved in text comprehension. Common research questions involve the existence and extent of parallel processing and the influence of lexical and syntactic factors on reading behaviour. This tradition commonly adopts well-constrained text processing, such as presenting a single sentence per screen. The text presented will conform to a clear design structure in order to pinpoint the exact mechanisms of oculomotor control during reading. Hence, 'reading' in the higher-level sense, such as literary comprehension of a novel, is not the impetus of the reading research tradition from an eye movement perspective. With higher-level reading, factors such as genre, education level, and discourse structure are the main predictors, as opposed to word frequency, word length, number of morphemes etc. in reading research on eye-movement control. The well-constrained nature of reading research, as well as consistent dedication within the field has generated a very well-researched domain where the level of sophistication is high. Common measures of interest to reading researchers are first fixation durations, first-pass durations and the number of between- and within-word regressions. Unique to reading research is the stimulus lay-out which has an inherent order of processing (word one comes before word two, which comes before word three...). This allows for measures which use order as a component, regressions for instance, where participants re-fixate an already fixated word from earlier in the sentence. Reading research has also spearheaded the use of gaze-contingent display changes in eye-tracking research. Here, words can be changed, replaced, or hidden from view depending on the current locus of fixation (e.g. the next word in a sentence may be occluded by (x)s, just delimiting the number of characters, until your eyes land on it, see page 50). Gaze-contingent eye tracking is a powerful technique to investigate preview benefits in reading and has been employed in other research areas to study attention independently from eye movements. Good overview or milestone articles in reading research are Reder (1973); Rayner (1998); Rayner and Pollatsek (1989); Inhoff and Radach (1998); Engbert, Longtin, and Kliegl (2002).
- *Scene perception* is concerned with how we look at visual scenes, typically presented on a computer monitor. Common research questions concern the extent to which various bottom-up or top-down factors explain where we direct our gaze in a scene, as well as how fast we can form a representation of the scene and recall it accurately. Since scenes are presented on a computer screen, researchers can directly manipulate and test low-level parameters such a luminance, colour, and contrast, as well as making detailed quantitative predictions from models. Typical measures are number of fixations and correlations between model-predicted and actual gaze locations. The scene may also be divided into areas of interest (AOIs), from which AOI measures and other eye

movement statistics can be calculated (see Chapter 6 and Part III of the book respectively). Suggested entry articles for scene perception are Henderson and Hollingworth (1999), Henderson (2003) and Itti and Koch (2001).

- *Usability* is a very broad research tradition that does not yet have established eye-tracking conventions as do the aforementioned traditions. However, usability research is interesting because it operates at a higher analysis level than the other research traditions, and is typically focused on actual real-world use of different artefacts and uses eye tracking as a means to get insight into higher-level cognitive processing. Stimulus and task are often given and cannot be manipulated to any larger extent. For instance, Fitts, Jones, and Milton (1950) recorded on military pilots during landing, which restricted possibilities of varying the layout in the cockpit or introducing manipulations that could cause failures. Usability is the most challenging eye-tracking research tradition as the error sources are numerous, and researchers still have to employ different methods to overcome these problems. One way is using eye tracking as an explorative measure, or as a way to record post-experiment cued retrospective verbalizations with the participants. Possible introductory articles are Van Gog, Paas, Van Merriënboer, and Witte (2005), Goldberg and Wichansky (2003), Jacob and Karn (2003), and Land (2006).

As noted, broad research traditions like those outlined above are often accompanied by specific experimental paradigms, set procedures which can be adapted and modified to tackle the research question in hand. We have already mentioned gaze-contingent research in reading, a technique that has become known as the *the moving-window paradigm* (McConkie & Rayner, 1975). This has also been adapted to study scene perception leading to Castelhano and Henderson (2007) developing the *flash-preview moving-window paradigm*. Here a scene is very briefly presented to participants (too fast to make eye movements) before subsequent scanning; the eye movements that follow when the scene is inspected are restricted by a fixation-dependent moving window. This paradigm allows researchers to unambiguously gauge what information from an initial scene glimpse guides the eyes.

The *Visual World Paradigm* (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) is another experimental set-up focused on spoken-language processing. It constitutes a bridge between language and eye movements in the 'real world'. In this paradigm, auditory linguistic information directs participants' gaze. As the auditory information unfolds over time, it is possible to establish at around which point in time enough information has been received to move the eyes accordingly with the intended target. Using systematic manipulations, this allows the researchers to understand the language processing system and explore the effects of different lexical, semantic, visual, and many other factors. For an introduction to this research tradition, please see Tanenhaus and Brown-Schmidt (2008) and Huettig, Rommers, and Meyer (2011) for a detailed review.

There are also a whole range of experimental paradigms to study oculomotor and saccade programming processes. The *anti-saccadic paradigm* (see Munoz and Everling (2004) and Everling and Fischer (1998)) involves an exogeneous attentional cue—a dot which the eyes are drawn to, but which must be inhibited and a saccade made in the *opposite* direction, known as an *anti-saccade*. Typically anti-saccade studies include more than just anti-saccades, but also pro-saccades (i.e. eye movements *towards* the abrupt dot onset), and switching between these tasks. This paradigm can therefore be used to test the ability of participants to assert executive cognitive control over eye movements. A handful of other well-specified 'off-the-shelf' experimental paradigms also exist, like the anti-saccadic task, to study occulomotor and saccade programming processes. These include but are not limited to: the *gap task* (Kingstone & Klein, 1993), the *remote distractor effect* (Walker, Deubel, Schneider, & Findlay, 1997),

saccadic mislocalization and *compression* (Ross, Morrone, & Burr, 1997). Full descriptions of all of these approaches is not within the scope of this chapter; the intention is to acquaint the reader with the idea that there are many predefined experimental paradigms which can be utilized and modified according to the thrust of your research.

3.2 What caused the effect? The need to understand what you are studying

A basic limitation in eye-tracking research is the following: it is impossible to tell from eye-tracking data alone what people think. The following quote from Hyrskykari, Ovaska, Majaranta, Räihä, and Lehtinen (2008) nicely exemplify how this limitation may affect the interpretation of data:

> For example, a prolonged gaze to some widget does not necessarily mean that the user does not understand the meaning of the widget. The user may just be pondering some aspect of the given task unrelated to the role of the widget on which the gaze happens to dwell. ... Similarly, a distinctive area on a heat map is often interpreted as meaning that the area was interesting. It attracted the user's attention, and therefore the information in that area is assumed to be known to the user. However, the opposite may be true: the area may have attracted the user's attention precisely because it was confusing and problematic, and the user did not understand the information presented.

Similarly, Triesch, Ballard, Hayhoe, and Sullivan (2003) show that in some situations participants can look straight at a task-relevant object, and still no working memory trace can be registered. Not only fixations are ambiguous. Holsanova, Holmberg, and Holmqvist (2008) point out that frequent saccades between text and images may reflect an interest in integrating the two modalities, but also difficulty in integrating them. That eye-movement data are non-trivial to analyse is further emphasized by the remarks from Underwood, Chapman, Berger, and Crundall (2003) which detail that about 20% of all non-fixated objects in their driving scenes were recalled by participants, and from Griffin and Spieler (2006) that people often speak about objects in a scene that were never fixated. Finally, Viviani (1990) provides an in-depth discussion about links between eye movements and higher cognitive processes.

In the authors' experience, it is very easy to get dazzled by eye-tracking visualizations such as scanpaths and heat maps, and assume for instance that the hot-spot area on a webpage was interesting to the participants, or that the words were difficult to understand, forgetting the many other reasons participants could have had for looking there. Its negative effect on our reasoning is known under the term 'affirming the consequent' or more colloquially 'backward reasoning' or 'reverse inference'.

We will exemplify the idea of backward reasoning using the music and reading study introduced on page 5. This study was designed to determine whether music disturbs the reading process or not. The reading process is measured using eye movements. These three components are illustrated schematically in Figure 3.1. In this figure, all the *(m)*s signify properties of the experimental set-up that were manipulated (e.g. the type of music, or the volume level). The *(c)*s in the figure represent different cognitive processes that may be influenced by the experimental manipulations. The *(b)*s, finally, are the different behavioural outcomes (the eye movements) of the cognitive processes. Note that we cannot measure the cognitive processes *directly* with eye tracking, but we try to capture them indirectly by making manipulations and measuring changes in the behaviour (eye movement measures).[11]

[11] See Poldrack, 2006 for an interesting discussion regarding reverse inference from the field of fMRI.

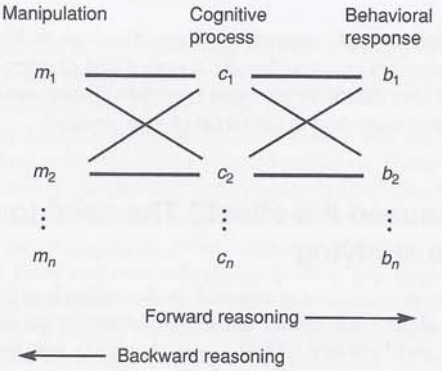


Fig. 3.1 Available reasoning paths: possible paths of influence that different variables can have. Our goal is to correctly establish what variables influence what. Notice that there is a near-infinite number of variables that influence, to a greater or lesser degree, any other given variable.

Each of the three components (the columns of Figure 3.1) introduce a risk of drawing an erroneous conclusion from the experimental results.

1. During data collection, perhaps the experiment leader unknowingly introduced a confound, something that co-occurred at the same time as the music. Perhaps the experiment leader tapped his finger to the rhythm of the music and disturbed the participant. This would yield the path $(m_2) \rightarrow (c_1) \rightarrow (b_1)$, with (m_2) being the finger tapping. As a consequence, we *do* get our result (b_1), falsely believing this effect has taken the path of $(m_1) \rightarrow (c_1) \rightarrow (b_1)$, while in fact it is was the finger tapping (m_2) that drove the entire effect.
2. We hope that our manipulation in stage one affects the correct cognitive process, in our case the reading comprehension system. However, it could well be that our manipulation evokes some other cognitive processes. Perhaps something in the music influenced the participant's confidence in his comprehension abilities, (c_2), making the participant less confident. This shows up as longer fixations and additional regressions to double-check the meaning of the words and constructions. Again, we do get our (b_1), but it has taken the route $(m_1) \rightarrow (c_2) \rightarrow (b_1)$, much like in the case with long dwell time on the widget mentioned previously.
3. Unfortunately, maybe there was an error when programming the analysis script, and the eye-movement measures were calculated in the wrong way. Therefore, we think we are getting a proper estimation of our gaze measures (b_1), but in reality we are getting numbers representing entirely different measures (b_2).

Erroneous conclusions can either be *false positives* or *false negatives*. A *false positive* is to erroneously accept the null hypothesis to be false (or an alternative explanation as correct). In Figure 3.1 above, the path $(m_1) \rightarrow (c_2) \rightarrow (b_1)$ would be such a case. We make sure we present the correct stimuli (m_1), and we find a difference in measurable outcomes (b_1), but the path of influence never involved our cognitive process of interest (c_1), but some other function (c_2). We thus erroneously accepted that (c_1) is involved in this process (or more correctly: falsely rejected that it had no effect). The other error is the *false negative*, where we erroneously reject an effect even though it is present and genuine. For example, we believe we test the path $(m_1) \rightarrow (c_1) \rightarrow (b_1)$, but in fact we unknowingly measure the wrong eye-movement variables (b_2) due to a programming error. Since we cannot find any differences

in what we believe our measures to be, we falsely conclude that either our manipulation (m_1) had no effect, or our believed cognitive process (c_1) was not involved at all, when in fact if we had properly recorded and analysed the right eye-movement measures we would have observed a significant result. False negatives are also highly likely when you have not recorded enough data; maybe you have too few trials per condition, or there are not enough participants included in your study. If this is the case your experiment does not have enough *statistical power* (p. 85) to yield a significant result, even though such an effect is true and would have been identified had more data been collected.

How can we deal with the complex situation of partly unknown factors and unpredicted causal chains that almost any experiment necessarily involves? There is an old joke that a good experimentalist needs to be a bit neurotic, looking for all the dangers to the experiment, also those that lurk below the immediate realm of our consciousness, waiting there for a chance to undermine the conclusion by introducing an alternative path to (b_1). It is simply necessary to constrain the number of possible paths, until only one inevitable conclusion remains, namely that: "(m_1) leads to (c_1) because we got (b_1) and we checked all the other possible paths to (b_1) and could exclude them". Only then does backward reasoning, from measurement to cognitive process, hold.

There is no definitive recipe for how to detect and constrain possible paths, but these are some tips:

- As part of your experimental design work, *list* all the alternative paths that you can think of. Brainstorming and mind-mapping are good tools for this job.
- Read previous research on the cognitive processes involved. Can studies already conducted exclude some of the paths for you?
- The simpler eye-movement measures belonging to fixations (pp. 377–389) and saccades (pp. 302–336) are relatively well-investigated indicators of cognitive processes (depending on the research field). The more complex measures used in usability and design studies are largely unvalidated, independent of field of research. We must recognize that without a theoretical foundation and validation research, a recorded gaze behaviour might indicate just about any cognitive process.
- If your study requires you to use complex, unvalidated measures, do not despair. New measures must be developed as new research frontiers open up (exemplified for instance by Dempere-Marco, Hu, Ellis, Hansell, & Yang, 2006; Goldberg & Kotval, 1999; Ponsoda, Scott, & Findlay, 1995; Choi, Mosley, & Stark, 1995; Mannan, Ruddock, & Wooding, 1995). This is necessary exploratory work, and you will have to argue convincingly that the new measure works for your specific case, and even then accept that further validation studies are needed.
- Select your stimuli and the task instructions so as to constrain the number of paths to (b_1). Reduce participant variation with respect to background knowledge, expectations, anxiety levels, etc. Start with a narrow and tightly controlled experiment with excellent statistical power. After you have found an effect, you might have to worry about whether it generalizes to all participant populations; is it likely to be true in all situations?
- Use *method triangulation*: simple additional measurements like retention tests, working memory tests, and reaction time tests can help reduce the number of paths. Hyrskykari *et al.* (2008), from whom the quotes above came, argue that retrospective gaze-path stimulated think-aloud protocols add needed information on thought processes related to scanpaths. If that is not enough, there is also the possibility to add other behavioural measurements. We will come back to this option later in this chapter (p. 95).

3.2.1 Correlation and causality: a matter of control

A fundamental tenet of any *experimental* study is the operationalization of the mental construct you wish to study, using *dependent* and *independent* variables. Independent variables are the causal requisites of an effect, the things we directly manipulate, $(m_i, i = 1, 2, \ldots, n)$ in Figure 3.1. Dependent variables are the events that change as a direct consequence of our manipulations—our independent variables are said to *affect* our dependent variables. This terminology can be confusing, but you will see it used a lot as you read scientific eye-tracking literature so it is important that you understand what it means, and the crucial difference between independent and dependent variables. In eye tracking your dependent variables are any of the eye-movement measures you choose to take (as extensively outlined in Part III).

A perfect experiment is one in which no factors systematically influence the dependent variable (e.g. fixation duration) other than the ones you control. The factors you control are typically controlled in groups, such as 'listens to music' versus 'listens to cafeteria noise' or along a continuous scale such as introversion/extroversion (e.g. between 1 and 7). A perfectly controlled experimental design is the ideal, because it is only with controlled experimental designs that we are able to make statements of causality. That means, if we manipulate one independent variable while keeping all other factors constant, then any resulting change in the dependent variable will be due to our manipulated factor, our independent variable (as it is the only one that has varied).

In reality, however, all experiments are less than perfect, simply because it is impossible to control for every single factor that could possibly influence the dependent variable. A correlational study allows included variables to vary freely, e.g. a participant reading to music could be influenced by the tempo of the songs, the genre, the lyrics, or simply the loudness of the music. If all these variables correlate with each other, it is not possible to separate the true influencing variable from the others. This results in the problem that we cannot know anything about the *causality* involved in our experiment. Perhaps one factor influences the dependent variable, or it could be that our dependent variable is actually causing the value of one of our 'independent' variables. Or, both variables could be determined by a third, hidden, variable. Lastly, they could be completely unrelated. Let us look at two examples from real life.

A psycholinguist wants to investigate the effect of prosody on visual attention. The experiment consists of showing pictures of arrays of objects while a speaker describes an event involving the objects. The auditory stimuli are systematically varied in such a way that one half of the scenes involve an object that is mentioned with prosodic emphasis, while the other half is not emphasized at all. A potentially confounding factor is the speaker making an audible inhale before any emphasis. This inhale is a signal to the participant to be on the alert for the next object mentioned, but it is not considered a prosodic part of the emphasis (which in this case includes only pitch and volume). In this example, the inhale *systematically* co-varies with the manipulated, independent variable, and may lead to false conclusions. Confounding factors may also co-vary in a random way with the independent variable. Such unsystematic co-variation is cancelled out given enough trials.

As another example, consider an educational psychologist testing the readability of difficult and easy articles in a newspaper. The hypothesis is that easier articles have a larger relative reading depth, because readers do not get tangled up with complex arguments and difficult words. A rater panel has judged different articles as being more or less difficult, on a 7-point scale. So we let the students read the real newspaper containing articles with both degrees of difficulty. Our results show that the easier articles have a larger reading depth. However, the readers are biased to spend more energy reading articles with interesting topics, and read them with less effort. Therefore, interesting articles have a lower difficulty rating.

This is our first hidden factor. Furthermore, the most interesting articles are placed early in the newspaper, which the reader attends to when most motivated. As the reader reads on, he skips more and more. Because of this, the least interesting articles (misjudged as difficult), are skipped to a larger extent—not because they are difficult, but because they correlate with late placement order, our second hidden factor. The net result is that we end up with an experiment purporting to show an effect due to difficulty/ease of articles, while the real effect is driven by interest and placement.

The bottom line is that it is impossible to control all factors, but with the most important factors identified, controlled, and systematically varied, we can confidently claim to have a sound experimental design. The first scenario is such, because the stimuli are directly manipulated to include almost all relevant factors. The second example is more tricky. We are biased by the panel of raters and trust them to provide an objective measurement of a predictor variable, but the raters are only human. Experiments such as these are also typically presented as experimental in their design, although they are much more sensitive to spurious correlations than our first scenario. The key problem is that the stimuli are not directly controlled. The newspaper has the design it has, and the articles are not presented in a random and systematically varied manner. By allowing important factors to covary, we end up with a design that is susceptible to correlations and is more likely to produce false conclusions.

It should nevertheless be remembered that increasing experimental control tends to decrease ecological validity and generalizability of the research. Land and Tatler (2009) in their preface express concern over the "passion for removing all trace of the natural environment from experiments" they see with many experimental psychologists. Accepting the loss of some control may often be a reasonable price to pay to be able to make an ecologically valid study. In the end, we do want to say something about performance in the real world. An example of the difference between the real world and the laboratory is presented by Wang *et al.* (2010), who found a greater number of dwells to in-car instruments during field driving compared to simulator driving.

3.2.2 What measures to select as dependent variables

Designing a study from scratch often involves the very concrete procedure of drawing the eye-movement behaviour you and your theories predict on print-outs of your stimuli, and matching the lines you draw with candidate measures from Chapters 10–13. Some of these measures are relatively simple, while others are complex. Often, you may inherit your measures from the paradigm you are working in, or the journal paper you are trying to replicate. Your study may also be so new that you need to employ rarely used, complex measures. After you have selected some measures, run a pilot recording and make a pilot analysis with those measures. In either case, you should strive to select your measures during the designing phase of the experiment, and make sure they work with your eye-tracker, the stimuli, your task, and the statistics you plan to use.

Note that as a beginning PhD student, you may have to spend up to a whole year until the experiment is successfully completed, but with enough experience the same process can be reduced to as little as a month. It is seldom the data recording experience, nor the theoretical experience that makes this difference, but the experience in how to design and analyse experiments with complex eye-movement measures. Making sure appropriate measures are used will certainly save you time during the analysis phase and possibly prevent you from having to redesign the experiment and record new data.

Frequently, the complex measures inherit properties of the simpler ones. For instance, transition matrices, scanpath lengths, and heat map activations depend on how fixations and saccades were calculated, which in turn depend on filters in the velocity calculation. It is not

straightforward to decide which measures to choose as dependent variables, as this choice depends on many different considerations.

In addition, complex measures such as transition diagrams (p. 193), position dispersion measures (p. 359), and scanpath similarity measures (p. 346) have not yet been subjected to validation tests, or used over a number of studies to show to which cognitive process they are linked. Active validation work exists only for a few simple measures from within scene perception, reading, and parts of the neurological eye-tracking research, for instance smooth pursuit gain (p. 450) and the anti-saccade measures (p. 305). These measures have been used extensively, and we have gathered considerable knowledge about what affects their values in one or the other direction.

An initial factor concerns the possibilities and the limitations of the hardware that you use. Animated stimuli, for instance, invalidate fixation data from all algorithms that do not support smooth pursuit detection (p. 168). Second, the sampling frequency (p. 29) may limit what measures you can confidently calculate. Third, the precision of the system and participants (p. 33) may exert a similar constraint. Fourth, relatively complex measures may require extensive programming skills or excessive manual work (in particular with head-mounted systems, p. 227), making them not a viable option for a study. Finally, some measures are more suitable for standard statistical analysis than others.

Therefore, in any type of eye-tracking project, part of the experimental design consists of selecting measures to be used for dependent variables, and to verify that the experimental set-up and equipment make it possible to calculate the measures. It is definitely advisable, in particular when using new experimental designs, to use the data collected in the pilot study (an essential check-point, described on p. 114) to verify that the method of analysis, including calculation of measures, actually works.

If you are at the start of your eye-tracking career, the approach of already thinking about the analysis stage when you are designing the experiment forces you to think through the experiment carefully and to design it so it answers your research question faster, more accurately, and with less effort. The eyes should always be on the research question, and eye-tracking is just a tool for answering it.

Question the *validity* and *reliability* of your measures. Validity is whether the dependent variable is measuring what you think it is measuring, for instance you may assume longer dwell time is a good index of processing difficulty in your experiment, but in fact this reflects preferential looking at incongruous elements of your stimulus display. Reliability refers to replicable effects; your chosen measure may give the same value over-and-over, in which case it is reliable, but note that a reliable measure is not necessarily a valid one (see page 463 for an extended discussion). You may find longer dwell times time and time again, which are not a measure of processing difficulty, as you thought, but rather a measure of incongruity. Below is a quick list on how to select your eye-tracking measures keeping the above issues in mind:

- Obviously, select the measure which fits your hypothesis best. If you think that your text manipulation will yield longer reading times, then first-pass duration or mean-fixation duration are likely measures, but number of regressions only an indirect (but likely correlated) measure. Unless of course your hypothesis is actually that reading times will be longer due to more regressions.
- Are you working within an established paradigm? Use whatever is used in your field to maximize the compatibility of your research.
- Identify other functionally equivalent measures for your research question. Are you interested in mental workload for example? Then find out what other measures are

used to investigate this, for instance using the index. Perhaps some of the alternative measures are better and completely missed by you and others in your paradigm.

- Prioritize measures that have been extensively tested, as there is better insight into potential factors affecting them. For example, first fixation durations in reading have been tested extensively and we know how they will react to changes in, for instance, word frequency. It would be less problematic to do a reverse inference with this kind of measure (using the first fixation durations to estimate the processing difficulty increase of a manipulation) than with other less well-explored measures.
- Select measures that are as fine-grained as possible, for example measures that focus on particular points in time rather than prolonged gaze sequences. This allows you to perform analyses where you identify points in time where the participant is engaged in the particular behaviour in which you are interested, e.g. searching behaviour, and then extract just the measures during just these points. This is more powerful than just extracting all instances of this measure during the whole trial, where the particular behaviour of interest is mixed with many other forms of gaze behaviour (which essentially just contribute noise to your results).
- To minimize problems during the statistical analysis, select measures that are either certain to generate normally distributed data, or measures that generate several instances per trial. In the latter case, if you cannot transform your data adequately or suffer from zero/null data, then you can take the mean of the measures inside the trials.[12] The implications of the central limit theorem are that a distribution of means will be normally distributed, regardless of the distribution of the underlying data. You will now have sacrificed some statistical power in order to have a well-formed data distribution which does not violate the criteria of your hypothesis tests. See Figure 3.2 for an example of different means to which you can aggregate. In this example, there is only one measurement value per trial, but repeated measurements within each trial would have provided even more data to either keep or over which to aggregate.

3.2.3 The task

Eye-tracking data is—as shown very early by Yarbus (1967, p. 174ff) and Buswell (1935, p. 136ff)—extremely sensitive to the task, so select it carefully. A good task should fulfil three criteria:

1. The task should be neutral with regard to the experimental and control conditions. The task should not favour any particular condition (unless used as such).
2. The task should be engaging. An engaging task distracts the participant from the fact that they are sitting in, or wearing, an eye-tracker and that you are measuring their behaviour.
3. The task should have a plausible cover story or be non-transparent to the participant. This stops the participant from second-guessing the nature of the experiment and trying to give the experimenter the answers that she wants. When the experiment itself causes the effects expected it is said to have *demand characteristics*.

If you are afraid to bias them, then give participants a very neutral task, but remember that weak and overly neutral tasks may also make each participant invent their own task. If you present an experiment with 48 trials and you do not provide a task, you are not to be surprised

[12] Note that if a level of averaging is severely skewed by outlying data points, it might be more appropriate to take a median at the trial or participant level.

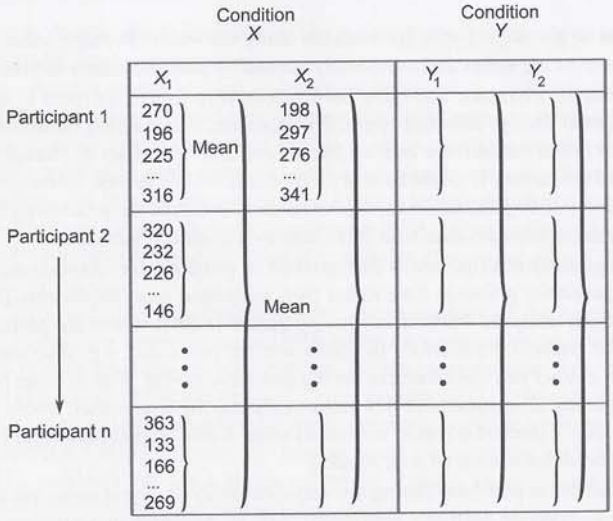


Fig. 3.2 A typical experimental design, and how means are calculated from it. Here we have two independent variables, X and Y, each with two factors, $X_{1,2}$ and $Y_{1,2}$. The conditions could be preferred and non-preferred music, each with high and low volume level, for instance. Each number represents an eye-movement measure from one trial.

if you find that the participants have been looking outside of the monitor, daydreaming, or falling asleep. Very general tasks such as "just look at the images" may require some mock questions to make the participants feel like they can provide answers/reactions to the stimuli. If you show pictures and want to make it probable that participants indeed scan the picture, a very neutral mock question could be "To what degree did you appreciate this image?". This question is neutral in the sense that it motivates participants to focus attention on the image presented, but still does not bias their gaze towards some particular part or object in the scene. Furthermore, if you add random elements, such as asking alternating questions and only at a random 30% of the trials, it reduces tediousness and predictability.

Tasks can also be used very actively in the experimental design, which was what Yarbus did, showing the same image to a participant but with differing instructions, thereby creating experimental conditions. In such a case the overt task starts and drives the experimental condition. A motivating task can also be the instruction to solve a mathematical problem, or to read so that the participants can answer questions afterwards. An engaging task can consume the full interest of the participants and surplus cognitive resources are aimed at more thoroughly solving the task. Additionally, an engaging task is not as exhausting for the participant, thus he can do more trials and provide you with more data. An important property of an engaging task is that it makes sense to the participant and allows him to contribute in a meaningful way.

In a general sense, the task starts when you contact potential participants, and talk to them about the experiment. When you recruit your participants, you must give them a good idea about what they are going to do in your experiment, but you should only tell them about the task you present to them. You should not reveal the scientific purpose of your study, since prior knowledge of what you want to study may make them behave differently. Suppose for instance that the researcher wants to show that people who listen to a scene description re-enact the scene with their eyes, as in Johansson, Holsanova, and Holmqvist

(2006). If a participant knows that the researcher wants to find this result, the participant is likely to think about it and to want to help, consciously or not, in obtaining this result, thus inflating the risk of a false positive. Such knowledge can be devastating to a study. For certain sensitive experiments, it may be necessary to include many distractor trials to simply confuse the participants about the hypotheses of the experiment. Additionally, our researcher should give participants a cover story, to be revealed at debriefing, that goes well with the kind of behaviour and performance she hopes participants will exhibit. For example:

> Throughout, participants were told that the experiment concerned pupil dilation during the retelling of descriptions held in memory. It was explained to them that we would be filming their eyes, but nothing was said about our knowing in which directions they were looking. They were asked to keep their eyes open so that we could film their pupils, and to look only at the white board in front of them so that varying light conditions beyond the board would not disturb the pupil dilation measurements (excerpt from Johansson *et al.* (2006), procedure section).

When you have settled on a task instruction that you feel fulfils the listed criteria sufficiently, then it is a good idea to write down the instructions. Written instructions allow you to give exactly the same task to all participants, rather than trying to remember the instructions by heart and possibly missing small but important parts of the task. Written instruction also help negate any *experimenter effects*: subtle and unconscious cues from the experimenter giving hints to the participant on how to perform.

3.2.4 Stimulus scene, and the areas of interest

Stimuli are of course selected according to the research question of the study in hand, and can be anything from abstract arrays of shapes or text, to scenes, web pages, movies, and even the events that unfold in real-world scenarios such as driving, sport, or supermarket shopping.

Scenes can roughly be divided up into two groups:

- Natural and unbalanced scenes, where objects are where they are and you do not control for their position, colour, shape, luminance etc. An example would be the real-world environment we interact with every day.
- Artificial and balanced scenes, which consist of objects selected and placed by the experimenter. For example, a scene constructed from clip arts, or a screen with collections of patches with different spatial frequency.

The two types offer their own benefits and drawbacks. Natural scenes, on average, will generalize better to the real world, as they are often a part of it or mimic it closely. If you find that consumers have a certain gaze pattern in a cluttered supermarket scene, you do not necessarily have to break down the scene into detailed features such as colour, shape, and contrast, but rather you can just accept that the gaze pattern works in this environment and not try to generalize outside of it. After all, the scene can be found naturally and this gaze pattern will at least work for this situation.

On the other hand, if you want to generalize across different scenes, you need a tighter control on all possible low-level features of the scene. This is where artificially constructed scenes work best, because you can manipulate the features and arrange them as you see fit.

In your efforts to control the scene, you should be aware of what attracts attention and consequently eye movements. This is especially challenging when you want to compare two types of natural scenes. Artificial scenes can be controlled on the detail level, but natural scenes usually cannot. If you want to compare two types of supermarket scenes to investigate which supermarket has the best product layout strategy, with varying products, it is impossible to completely control every low-level feature of the scenes. You just have to accept that

colour, luminance, contrast, etc. vary, and try to set up the task so the layout strategy will be the larger effect which drives your results. Perhaps, you can add low-level features post-hoc as covariates in the analysis, by extracting them from the scene video, to at least account for their effect.

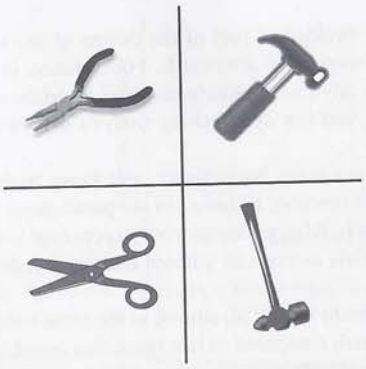
When selecting the precise stimuli, it is useful to consider what it is that generally draws our attention, so the effect of primary interest is not blocked or completely dominated by other larger effects. Below are a few examples of factors that are known to influence the allocation of visual attention, more can be found on pages 394–398:

- People and faces invariably draw the eyes, so if you want to study what vegetation elements in a park capture attention, you should perhaps not include people or evidence of human activity in the stimulus photos.
- If you use a monitor, the participants are likely to look more at the centre than towards the edges. They are also more likely to make more horizontal than vertical saccades, and very few oblique ones.
- Motion is likely to bring about reflexive eye movements towards it, irrespective of what is moving. Consider this if you want to conclude, for example, that bicycles capture drivers' attention more than pedestrians do; this may simply be because bicycles move faster, and nothing more.
- If you are looking at small differences in fixation duration, it matters whether you put stimuli in the middle or close to the edges of the monitor, because precision of samples will be lower at the extremities of the screen. The imprecision may force a premature end of the fixation by the fixation detection algorithm, and consequently cause your effect.
- Keep the brightness of your stimuli at approximately the same level, and also similar to the brightness of the calibration screen, or you may reduce data quality, as the calibration and measurement are performed on pupils with different sizes.

Stimulus images are often divided into AOIs, the 'areas of interest', which are sometimes also called 'regions of interest'. How to make this division is discussed in depth in Chapter 6. In short, the researcher chooses AOIs while inspecting potential stimulus pictures with the precise hypothesis and measures of the study in mind. Selecting AOIs while reviewing your already recorded data is methodologically dubious, because you may intentionally or subconsciously select your AOIs so that your hypothesis is validated (or invalidated). If you want to analyse what regions in your picture or film attracted participant gazes, but have the regions defined by the recorded data, you should use heat/attention map analysis (Chapter 7) rather than AOI analysis (Chapter 6).

As a very simple example of AOIs, you could show several pictures each with a matrix of objects in them, as in Figure 3.3(a), and determine whether visually similar items have more eye movements between them, than visually dissimilar objects. Simply construct an AOI around each object, and compare how many movements across categories versus within categories occur. Most eye-tracking analysis softwares allow for manual definition of AOIs as rectangles, ellipses, or polygons. AOIs are typically given names like 'SHEARS' and 'HAMMERS' to help keep track of the groups in the experimental conditions. This is also a perfect example of a case where it is easy to define the AOIs before the data recording, which should always be the preferred way.

When using film or animations as stimuli, as in Figure 3.3(b), where there are many moving objects, static AOIs are often of little use. Dynamic AOIs instead follow the form, size, and position of the objects as they move, which makes the data analysis easier. To the authors' knowledge, the first commercial implementation of dynamic areas of interest was



(a) Large square areas of interests with clear margins to compensate for minor offsets in data samples.

(b) Elliptical, but cropped, area of interest around the squirrel and the bird in the stimulus picture. Reproduced here with permission from the Blender foundation `www.bigbuckbunny.org`.

Fig. 3.3 Examples of AOIs.

made available in 2008, decades after the static AOIs began to be used. Dynamic AOIs come with their own set of methodological issues, however (p. 209).

3.2.5 Trials and their durations

A trial is a small, and most often, self-repeating building block of an experiment. In a minimal within-subjects design there may be as few as two trials in an experiment, for instance one trial in which participants look at a picture while listening to music, and another trial where they look at the picture in silence. The research question could be how music influences viewing behaviour. Or there may be several hundreds of trials in an experiment, for instance pictures of two men and two women, with varying facial expressions, hair colour, types of clothes, eye contact, etc., to see if those properties influence participants' eye movements.

In an experimental design, trials are commonly separated in time by a central fixation cross. For instance, you may have an experiment in which you first show a fixation cross in the middle of the screen, then remove the cross so as to have a blank screen while at the same time playing the word 'future' auditorily. The crucial period of time for eye-movement recording here is the blank screen, but it could equally be an endogenous spatial cue to the left or right or some other manipulation. The idea behind this experiment would be to see if time-related words such as 'future' or 'past' make participants look in specific directions. The trial sequence (fixation cross, stimulus presentation, and so on) will then iterate until the specified number of trials corresponding to that condition of the experiment has been fulfilled. Experimental trials are often more complex than this however, and may contain features like varying *stimulus onset asynchrony*, where the flow of stimulus presentation during the trial is varied according to specified time intervals. Taking the *flash-preview moving-window paradigm* outlined above as an example, the brief length of time for which the first scene picture is displayed, before the following gaze-contingent display, can be varied corresponding to different durations. Võ and Henderson (2010) have implemented such a manipulation to shed light on just how much of a glimpse is necessary to subsequently guide the eyes in scene viewing, and they found that 50–75 ms is sufficient.

Thus, the 'layout' of a trial is usually decided as part of the design of the experiment. In some cases, however, trials must be reconstructed afterwards. For instance, in a study of natural speech production, a trial may start anytime a participant utters a certain word. Post-recording trials are more difficult to create, and few eye-tracking analysis software packages have good support for them.

Typically, an eye-tracking study involves many participants, and many trials in which different stimuli are presented. It is not uncommon, to have say 40 participants looking at 25 pictures with a duration of 5 seconds each. Always design your experiment to extract the maximum amount of data. Add as many trials as you can without making it tedious for the participants.

Moreover, we do not want all participants to look at all stimuli in the same order, because then they may look differently at early stimuli compared to late ones; this could be due to a *learning effect* or an *order effect*. The former case refers to when participants have become better at the task towards the end of the experiment, the latter case is when there is something about the order of presentation which biases responses and eye-movement behaviour. To avoid such confounds trial presentation is randomized for all 40 of your participants, no two participants viewing the stimuli in the same order. Then, any effects of learning or order will be evenly spread out across all stimulus images, and will not interfere with the actual effect that we want to study. Presentation order can usually be randomized by the experimental software. This is easy and usually enough to eliminate learning/order effects. Otherwise, a separate distinct stimulus order is prepared for each participant beforehand. This takes more time, but is virtually foolproof as it can be counter-balanced and randomized with a higher degree of control.

An old problem with scrambling stimulus presentation order, was that in your data files the first 5 seconds of each participant were recorded from different trials. It is not possible to place the first 5 seconds of data next to one another, participant by participant, as you would typically want to do when you calculate the statistical results comparing 20 of your participants to the other 20. Until very recently, eye-tracking researchers had to unscramble the data files manually, or write their own piece of software to do it for them. This was a very time-consuming and rather error-prone way to work with the data. Today, most eye movement recording software communicates with the stimulus presentation program so as to record a reference to the presented stimulus (such as the picture file name) into the correct position in the eye movement data file. Thus, the information for how to *de*randomize is in the data file, and can be used by the analysis software. Some eye-tracking analysis programs today allow users to derandomize data files fairly automatically, immediately connecting the right portion of eye-tracking data to the correct stimulus image, which simplifies the analysis process a great deal.

When showing sequences of still images that are all presented at a constant duration, participants may learn how much time they have for inspection and adopt search strategies that are optimized for the constant presentation duration (represented by thoughts such as, for instance "I can look up here for a while, because I still have time to look at the bottom later"). If such strategies undermine the study, *randomized variable trial durations* can be used to reduce predictability and counteract the development of visual strategies (see also Tatler, Baddeley, & Gilchrist, 2005).

Precise synchronization between stimulus onset and start of data recording for a trial is very important. Many factors may disturb synchronization and cause latencies that make your data difficult to work with or your results incorrect (p. 43). One potential problem is the loading time of stimulus pictures in your stimulus presentation program. If for some reason you show large uncompressed images (e.g. large bitmaps) as stimuli, and send the start of a recording signal just before presenting the picture, the load time of the picture until it is

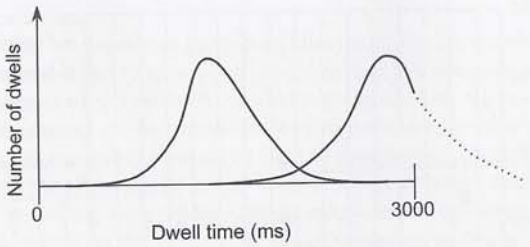


Fig. 3.4 Dotted line of one distribution of dwell times is outside of the fixed trial duration.

shown may be in the order of hundreds of milliseconds, which means that your participants do the first saccades and fixations not on the picture (which you will think when looking at data), but on the screen you showed before the picture was loaded. The solution to the loading problem is to pre-load images into memory before they are shown. Playing videos for stimuli requires an even more careful testing of synchronization. Additionally, synchronizing the eye-tracker start signal with the screen refresh is important to avoid latencies due to screen updates, especially when using newer but slower flat-screen monitors, which typically operate at 60 Hz. Ideally, these issues should be taken care of by your particular stimulus presentation package, and these low-level timing issues are beyond the scope of this book.

Fixed trial durations in combination with a small number of AOIs may complicate variance-based statistical analysis for a number of measures, for instance dwell time (p. 386), reading depth (p. 390), and proportion over time analysis (p. 197). Figure 3.4 shows the distribution of dwell time on a single AOI presented in two different conditions measured in trials of 3000 ms length. In one of the two conditions, the distribution nicely centres around 1500 ms, and both tails are within bounds. In the second condition, however, the top of the distribution is close to the 3000 ms limit that part of what would have been its right tail has been cut off by the time limit of the trial.

3.2.6 How to deal with participant variation

In the planning stage, participants appear as abstract entities with very little or no individual variation or personal traits. Later, during recording, real people come to the laboratory and fill the abstract entities with what they are and do. It is important to see participants as both. In this section, 'participants' refers only to the abstract entities that provide us with data points, while on pages 115–116 we discuss participants as people.

A large proportion of the eye-tracking measures that have been examined have proven to be *idiosyncratic*, which means that every participant has his or her own basic setting for the value. Fixation duration, one of the most central eye-tracking measures, is idiosyncratic. In Figure 3.5, participants 2, 9, and 21 have long individual fixation durations, while participants 6 and 12 have short individual fixation durations. This is like their baseline. The figure shows that the variation between participants is much larger than it is within the participants. The difference between trials completely drowns in these idiosyncratic durations, and it means that we are actually trying to find a small effect within a much larger effect.

How can we deal with participant variability and idiosyncracy? Participants can be divided into groups and assigned tasks/stimuli in a variety of different ways. The two most common, used here only for exemplification, are the *within-* and the *between-subjects designs*. Table 3.1 shows these two varieties in our example with the four sound conditions. In a between-subjects design, the participants only read a text under one sound condition, either

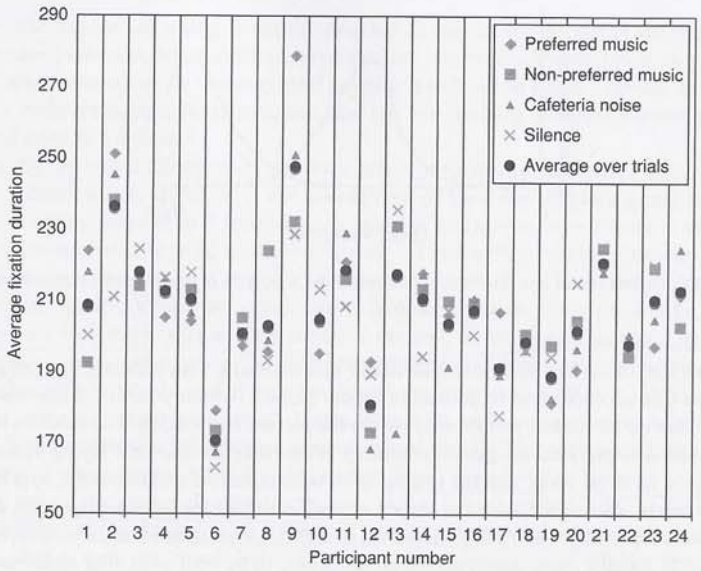


Fig. 3.5 Idiosyncracy: every participant has his or her own individual average fixation duration and exhibits it across different recordings. Individual participant variation is large, and the effect of the experimental conditions is small.

Table 3.1 Between- and within-subjects design in a task with four conditions (different sounds being played, or silence). S1 to S16 are the different participants. In the between-subjects design everyone reads one text to one type of sound, and then leaves the lab. In the within-subjects design, every participant has to read in all four sound conditions.

Condition	Between				Within			
Preferred music	S1	S5	S9	S13	S1	S2	S3	S4
Non-preferred music	S2	S6	S10	S14	S1	S2	S3	S4
Cafeteria noise	S3	S7	S11	S15	S1	S2	S3	S4
Silence	S4	S8	S12	S16	S1	S2	S3	S4

listening to music liked, music disliked, noise, or silence. So when we compare preferred music to unpreferred, we also compare two different participants to one another. In a within-subjects design, on the other hand, each participant reads texts in all four conditions, which means that a comparison between sound conditions is made within the same participant. It does require every participant to read four texts, which takes longer and may introduce learning effects (the last text is read differently than the first), which forces us to randomize. The within-subjects design also means that we must find four comparable texts so the effects we find are not driven by text differences rather than the investigated sound conditions.

In most psychological research, the effects sought after are usually so small that we need many trials to find them, making a within-subjects design the only practically available solution. Furthermore, this approach also lets us see to what extent the effect is representative in a larger population of participants. In a within-subjects design, since we try to find the effect for each individual, we can also see how many of the participants display the sought-after effect. If all participants display it, then the effect is highly generalizable to a larger population.

However, this does not mean that participant idiosyncracy is not a problem for your data. Unexplained variance still shows up as noise in your models and your ultimate goal is often to provide as full an explanation as possible of what is going on. This also means explaining or at least reducing the impact of idiosyncratic factors so you can more clearly see the effect of your manipulation and accurately estimate its size (data analysis programs like SPSS give you the option to output the *effect size* of a significant result you obtain). Statistical approaches such as multilevel modelling are good for adding random factors (participants and items) and modelling them in order to explain their effect and contribution to the variance, for example by using random intercepts and slopes for every participant or item in the regression. Nevertheless, as a rule of thumb, it is a good idea to reduce the heterogeneity of your participants if you want to establish that your manipulations have a statistically significant effect on their performance. Both the task and the reception of the participants into the laboratory can be used for this.

So are there any benefits to using a between-subjects design? Yes, but they depend on the experiment in hand. Any within-subjects design has some problem of potentially allowing the participants to guess the manipulation. Given enough trials, the participant notices the pattern, e.g. the presentation of common words versus unusual words, and starts guessing the nature of the experiment. Once he has figured out the aim of the experiment, the participant is very likely to behave as expected to please the experimenter. This can be solved by introducing filler trials to throw the participant off his hypotheses, but for very sensitive experiments it will be best to use a between-subjects design.

Furthermore, consider an experiment where we test the impact of two different instructions on problem solving. We give participants a problem to solve and provide them with one type of information. We cannot then present them with the same problem again and supply them with another type of information, as they carry the experience from the first instance with them. In other words, we only have one try per participant, and we have to use a between-subjects design. Given enough participants, we will be able to tell whether one type of information had a larger impact than the other type.

Naturally, if we use participants that are part of a fixed category, for instance dyslexics, then we are forced to use a between-subjects design as we can never 'switch-off' the dyxlexia for a participant and use him as his own baseline. Pre-occuring variables like this, which exist in your participants and you cannot directly manipulate, are known as *quasi* independent variables.

3.2.7 Participant sample size

Only when you know the experimental design can you estimate the number of participants you need for your study, but even then it requires an estimation of the variance in the data you have not yet collected.

It is often the case that the journal in which you plan to publish requires that each condition has a sufficient number of participants, for instance 10 contributing to each cell mean, or maybe more. If you have a 2-by-2 design, as in Figure 3.2—two independent variables, each with two levels—-you have four basic cells. In this figure the design structure is entirely within-subjects, but could equally be between-subjects, with different people listening to preferred or non-preferred music.[13] If this is the case, as different individuals relate to different participant groups, we would need more participants in total to achieve the minimum requirement of a sample size of 10 for each cell mean. The reason for a minimum sample

[13] In this particular case, we would actually have a *mixed design* here, music type would be manipulated between participants, while volume level would be manipulated within subjects.

size per cell is that we want to make sure that we have used enough data so that we do not prematurely dismiss our results as null. More participants and trials prevents us from making this mistake, giving us better *statistical power*. Failure to find a significant effect due to too low power, even though an effect is present, is what statisticians call a Type II error—a false negative.

It is also worth bearing in mind that you might lose participants along the way. It is common practice to exclude participants from the analysis of eye-tracking data due to poor data quality of the recording, or perhaps they simply did not do the task properly because they struggled to fully understand the task instructions, in which case they should also be removed. Insufficient data due to sample attrition is an issue which we will also address when we come to data recording in the next chapter.

Conversely, there is such a thing as too much data, as well as too little. Consider an experiment where we let participants read two types of text, one technical and one more casual, and measure the average fixation duration. With 20 participants in each of the two cells, we find significant differences at the $p < 0.05$ level. If we instead record 500 participant in each cell, we will very likely find that the signal-to-noise ratio has been amplified so the test is now significant at a $p < 0.001$ level. In other words, the probability that our observed differences in fixation durations between the technical texts and casual texts are due to chance is 1 in 1000. More data has made our result stronger, but it was not necessary to record data from so many participants.

Caution is needed with regard to large sample sizes, therefore, as it is potentially possible to find positive effects in almost any experimental manipulation you do. With enough data, any effect, however trivial, will cut through the random noise. Now, consider an alternative experiment where we again use 20 participants per cell, but do not find the expected effect of our manipulation. If we keep recording until we have 500 participants per cell, and we then observe a significant effect at the $p < 0.05$ level, we now run the risk of a an error similar to, but not quite, a Type I error—a false positive. Given enough data, small effects will be amplified until they qualify as significant. For example, during our manipulation, we happened to pick two texts which had slight and barely visible differences in the font type. With enough data, we found significant effects, not in our intended manipulation of text genre, but rather in the type of font used. We risk falsely assuming an effect of text genre when in fact there is none (but an effect of font type).

The optimal number of participants to use varies, but there are various approaches to solve this. One way would be to follow the canonical research in your particular research field and journals, and just use the same number of participants and items. If you believe your effect size will deviate from previous research, then take earlier studies and calculate their statistical power (what is called the *retrospective power*). You can then use this power value together with the expected magnitude of your effect to generate the required number of participants needed for each cell. There is software for doing power calculations, but they still require an educated guess of the effect of magnitude and its variance. When the result of our hypothesis test is null, high statistical power allows us to conclude with greater confidence that this result is genuine, and that it is very unlikely that an effect of the hypothesized magnitude or larger was present.

Often, we accept a risk of a type II error (known as β) which is larger than the risk of a type I error (α), because the former can require large amounts of data to negate, which is not feasible in a standard eye-tracking experiment. The risk we take entails ending up with results that falsely show no effect of our manipulation. This is deemed less problematic than type I errors. This is not to say, however, that type I errors, i.e. spurious and invalid effects, do not show up in eye-tracking data. This probably happens all the time, but they only really pose a threat to the research tradition if they are not understood by the researcher, not

questioned by the reviewer, or not replicated by the research community. The error can be one or a combination of many aspects of the experiment: poor precision and accuracy of the eye-tracking hardware, bad operationalizations of the mental construct and selection of dependent and independent variables, questionable synchronization between stimulus presentation and eye movement recordings. It is up to the researcher to decide whether it is more important to be confident that the effect is present, or if it is more important to be confident that the effect is not there. Statistical power is seldom reported as we are typically interested in positive effects and there is a publication bias for these effects. We should keep in mind though, that (failed) replications can be very interesting and then power becomes an important issue to correctly falsify previous findings.

It is beyond the scope of this book to discuss detailed power calculations, but two simple examples can be given to put power and sample size into perspective. These examples were calculated using the formulae and tables in Howell (2007) for simple one-way ANOVAs.

- If we want an α of 0.05 (we correctly accept 95% of all true effects) and a power of 0.80 (we correctly reject 80% of all false effects), then we need a sample size of 72 participants per cell (i.e. per experimental condition).
- Given an α of 0.05 and a power of 0.95, then we need a sample size of 119 per cell.

However, there is more to the discussion than just getting your results significant. For example, earlier studies may have just very few participants (Noton & Stark, 1971a: two and four participants; Gullberg & Holmqvist, 1999: five participant pairs), and even though the results may be significant, there is also the problem of generalizability. With four participants, it is likely that these people will deviate from the average person we want to generalize to. Typically, the hypothesis tests tell us how likely it is that a sample is drawn from a particular population or not. This assumes that the participants are randomly sampled from the population at large. In practice, this is never the case. It is a fact that the vast majority of academic research is carried out on university students; this is also true of eye tracking. Unfortunately, we cannot see that anybody will go through the challenge of doing completely randomized sampling of the population during the recruitment of participants to an experiment. We can only hope to be humble when drawing conclusions and making generalizations. However, a study with only four participants may still be interesting. Not because we can generalize from it (which we cannot), but because it may generate interesting hypotheses that we may proceed later to test with a full experiment. The point is to not present a case study as a full generalizable experiment, or vice versa.

3.3 Planning for statistical success

Once the data of your experiment have been collected, you will have one or several files with the raw data samples. At this point in the future, you should already have a clear idea what to do with this data. Typically, the subsequent analysis consists of four main steps, each of which is described in the following subsections.

3.3.1 Data exploration

Data exploration is not often discussed in textbooks, but is nevertheless an important part of the analysis. The main purpose of data exploration is to get to know the data in order to be able to account for choices that are made in later stages of the analysis. A secondary purpose, which is nevertheless also vital, is to check for possible errors in the data. It happens all too easily that data were coded erroneously or incorrectly measured when the experiment

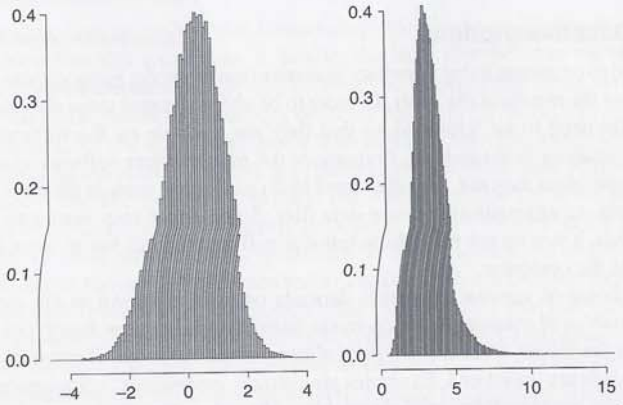
was carried out. Feeding the data into a data analysis without checking for errors may have devastating effects, either producing significant effects that do not exist, or hiding them.

The first goal of data exploration is to *check whether data quality is sufficient*. This can mostly be done in manufacturer software by inspecting the recorded data of individual participants. Position- and velocity-over-time diagrams, scanpath plots, and heat map visualizations are excellent tools to quickly inspect and judge the quality of data. For participants and trials who pass through this initial test, use event detection, AOI analysis, or the other methods in Chapters 5–9 to calculate values to those eye-movement measures that you have selected as so-called variables in your experiment.

Another main goal is to *look at the distribution of these variables*. A regular requirement for statistical tests is that the data are normally distributed (i.e. symmetrically distributed around the mean with values close to the mean being more frequent than values further away from the mean, compare the left part of Figure 3.6). As will become apparent in Part III of this book, many eye-tracking measures are not normally distributed. Eye-tracking measures, including fixation duration and most saccade measures, tend to have skewed distributions so that one tail of a histogram is thicker than the other tail, examplified in the right part of Figure 3.6. Skewed variables may become normally distributed after transformation, for instance, by computing the logarithm of the values, which may be the single most used transformation available. This transformation makes a positively skewed (typically right-skewed) distribution normal-looking by reducing higher values more than lower values. A distribution commonly log-transformed is human reaction time values, where there is a physical limit to how fast a human can respond to a stimulus, but no limit to how slow they can be. Therefore, the distribution typically has a fat positive tail consisting of the trials where the participant was fatigued, inattentive, or disrupted. A less common, but theoretically more powerful approach, is to analyse skewed distributions directly using methods developed for gamma distributions (if the untransformed values resemble this distribution). If the dependent variable is a proportion, especially outside the 0.3–0.7 range, then a log odds (logit) transformation is common. Navigating between transformations and methods for particular distributions becomes important during the analysis stage, especially so if you have limited data and cannot afford to aggregate it to produce a Gaussian distribution.

A third goal of exploratory data analysis is to *identify outliers*, that is, values that fall outside the normal range of measurements. These values need to be handled with care, as they may exert a disproportionately large influence on the results of the final analysis. Outliers may be the consequence of errors in the data recording or the event detection, or they may be actual rare measurements. In case they are errors, they need to be corrected or excluded. In case they are rare measurements, you may decide to leave them in or to exclude them. There are no strict guidelines about what to do with outliers. In some cases, it may be possible to predefine outliers. For instance based on previous experience and other research, one excludes all values that fall outside the range that is normally to be expected. In other cases this might not be possible and you need to decide which values are to be left out and which ones may stay in. This decision should ideally be made before the analysis is done. One strategy is to examine standardized values, and exclude values that are more than 3.29 standard deviations above or below the mean (Tabachnick & Fidell, 2000). Such rare values are not outliers by definition, however, since a few such extreme values are to be expected if the datafile is sufficiently large. Outliers may, finally, disappear spontaneously as a consequence of data transformation.

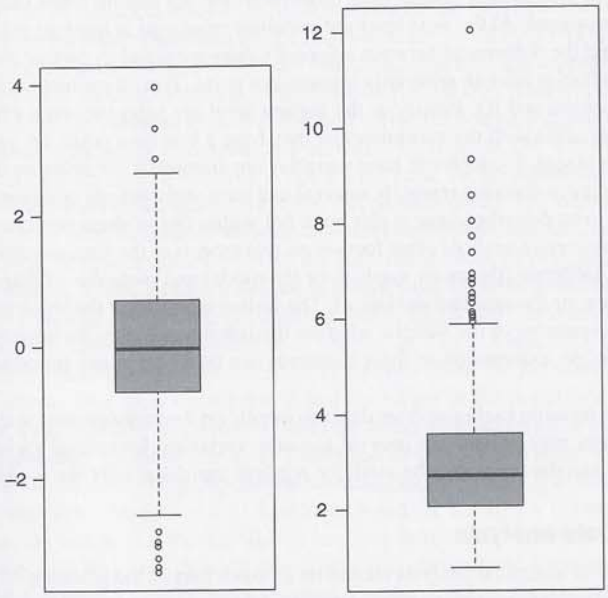
Plotting is also an indispensable tool in the later stages of data exploration. Particularly useful are box-and-whiskers plots, which give simultaneous information about the distribution as well as potential outliers (compare Figure 3.7). Additional plots that might be helpful are histograms (as in Figure 3.6), scatterplots, stem-and-leaf plots. In this stage of



(a) Normally distributed random variable.

(b) Positively skewed random variable.

Fig. 3.6 Histograms, symmetric and skewed, respectively.



(a) Normally distributed random variable.

(b) Positively skewed random variable.

Fig. 3.7 Boxplots of the variables shown in Figure 3.6.

the analysis, it is wise to make a plot for each participant separately as well as for each item. In that way, it becomes possible to identify potentially deviant participants or items that need to be excluded from further analysis.

3.3.2 Data description

Data description means using summary statistics (mean, mode, variance, etc.) to present in a concise way the results of the study. In order to be able to present these statistics, the available data usually need to be formatted so that they are readable by the software package with which the analysis is carried out. Sometimes the manufacturer software can do part of this job, but more often than not, you may need to do additional work in the form of transposing, restructuring, or aggregating the raw data files. Since errors may steal into the data at this stage as well, it is wise not to do these transformations by hand, but to leave them as much as possible to the computer.

The choice of summary statistics depends on what is known as *the measurement scale* of the variables of interest. An often-made distinction is between four types of measurement scales. At the lowest level are *categorical* or *nominal* variables. These take different values, but the values are unordered. Examples are colours, professions, grammatical categories, and so on. At the next level are *ordinal* variables. The values that these can take may be ordered, but the differences between adjacent values need not be the same. An example is the order in which a participant looks at different AOIs in an image. The participant may, by way of illustration, first look for a long while at one AOI, and then only briefly at the next before going on to a third AOI. These time differences are not visible when only the order of the AOIs is measured. At the next level are variables measured at *interval* scale. Values may be ordered and the differences between adjacent values are equal. A further characteristic is that interval variables have an arbitrarily chosen zero point. Typical examples of interval variables are temperature and IQ. Finally, at the highest level are *ratio* variables which are similar to interval variables with the exception that they have a true zero point, i.e. zero means that the variable is absent. Examples of ratio variables are dimension variables such as height, width, and time. In eye-tracking research, interval and ratio variables are common, and many of the measures to be described later in this book fall within one of these two categories.

The descriptive analysis often focuses on two aspects of the data, usually termed measures of central tendency (the mean, median, or the mode) and measures of dispersion (the range, the variance, or the standard deviation). The former summarize the value that is in a way the most representative of the sample, whereas the latter summarize the amount of variability in the sample. An explanation of these measures can be found in any introductory textbook on statistics.

Which measure to choose from depends largely on the measurement scale of the variables. All measures may be used for interval and ratio variables; for ordinal variables, the median, the mode, and the range may be used; for nominal variables, only the mode may be used.

3.3.3 Data analysis

The choice of statistical analysis should be as much part of the planning of a study as any of the other considerations given in this chapter. Statistical tests cannot be adapted so that they fit any kind of experimental design. Rather, the design of the study needs to be adapted so that the data can be analysed by an existing statistical test. If the choice of the test is not taken into account during the planning stages of the study, there is a risk that the results cannot be analysed properly, and, consequently, that drastic data transformations severely reduce the statistical power or, ultimately, that all the effort that was taken to run the study has been in vain.

The principle behind statistical testing is the following. The participants (and the materials) constitute a sample that is taken from some population of interest, for example, normal-reading adults, dyslectic children, second-language learners, and so on. A population is usually large, making it impossible to measure all of its members. The sample, thus,

is a non-perfect image of reality and consequently there is some degree of uncertainty in the results. Note that this uncertainty is smaller for large samples than for small samples. This uncertainty is also known as 'sampling error'. Sampling error is the variability that is for instance the consequence of measuring different participants, or the same participants on different occasions, or the same participants with different stimulus material (see also page 83). The purpose of inferential statistics is to distinguish sampling error from variability that may be related to another variable of interest. The outcome of the test is the probability that observed variability in the data is sampling error only. This probability is the *p*-value that is reported as the result of the test. If this probability is very low, then the conclusion is drawn that the variability in the data may be ascribed to variability in one or more variables.

During the past few decades, the possibilities for statistical analysis have greatly increased. There is now a large variety of different types of analysis available, some of which are simple, others more complex. The complex analyses are not necessarily better than the simple ones. A well-defined research question may be simple, and the accompanying analysis may be also. Perhaps the most important factor that determines the choice of the statistical analysis, and with that the design of the study, is that you select a test that you are comfortable with. As stated above, it is easier to adopt the design of an experiment to an existing statistical analysis then the other way around.

Different types of statistical analysis exist, depending on the variables that are included in the study. A rough two-way distinction can be made between parametric and non-parametric tests. Non-parametric tests (such as Wilcoxon, Friedman, sign test) are appropriate when the underlying dependent variable is ordinal or nominal. In eye-tracking research, ordinal dependent variables are not as common as interval or ratio variables. Nominal dependent variables, on the other hand, may occur frequently (for instance different AOIs). The distinction between an ordinal and an interval variable is not always clear. A three-point scale (e.g. cold-warm-hot) is without doubt an ordinal variable, but as more points are added to the scale, it increasingly resembles an interval variable. Nominal dependent variables are notoriously difficult to analyse. Simple statistical tests for the association between two nominal variables exist (e.g. chi-square, Fisher's exact test), but in practice the situation is usually more complicated. An overview of non-parametric tests is given in Siegel and Castellan (1988).

If the dependent variable is measured at an interval or a ratio scale, the statistical test is a parametric test. These tests rely on specific assumptions about the population from which the sample is drawn. One such assumption is that the values in the population are normally distributed, i.e. symmetrically distributed around the mean with values close to the mean being more frequent than values further away from the mean. Whenever there is evidence that the distribution of the underlying population is not normal there is a risk that the outcome of the test is unreliable. An option is to transform, using for instance a log or a square root transformation, the data so that the distribution becomes normal. The decision whether or not to transform the data may be a difficult one. There is a cost–benefit argument. The advantage is that the test results may be more reliable. The drawback is that the test results may become difficult to interpret as well as a loss of power. We lose power because, e.g. a log-transformation reduces large numbers more than small numbers, so we are less able to separate the difference between two large numbers. An alternative solution, which unfortunately is not ideal either, is to convert the measurement scale from interval/ratio to ordinal/nominal, and to do a non-parametric test. This solution is not ideal because this conversion involves loss of information, and with that loss of statistical power. We lose power if we ignore the size of the numbers and only focus on the sign (positive/negative), because we cannot distinguish between -1 and -100.

A different two-way distinction is whether there is one or several dependent variables. The collected history of eye-movement research give you access to much more than a single

measure for your study. When you have multiple dependent variables, you may decide to analyse them separately to see which of them yields significant differences between experimental groups. In doing so, the character of your study becomes exploratory rather than confirming or rejecting hypotheses. An alternative is to 'reverse the roles' of independent and dependent variables, and to see which of the dependent variables best predicts group membership. Suppose, for instance, that two experimental groups were involved in a study, for instance dyslexic readers and normal readers. These two groups all read a text and several measures are obtained from their reading: first fixation durations, number of inword regressions, gaze duration, saccadic amplitudes, etc. These measures can then be used as predictors to evaluate which of them predict whether a reader was a dyslexic or a normal reader. Finally, a number of multivariate statistical methods exist that may be used to see which variables 'group' together (for instance, factor analysis, principal component analysis, cluster analysis, correspondence analysis). This approach is exploratory rather than confirmatory. For an overview of different multivariate statistical analyses, we refer to Tabachnick and Fidell (2000).

Further factors that determine the choice of statistical analysis are the number and types of independent variables. In the following, we briefly describe a few types of analyses that are common within eye-tracking research. For each analysis, we provide a short example, and one or two references for further reading.

Analysis of variance or ANOVA is the appropriate analysis if the dependent variable is measured at interval or ratio scale and there are one or more independent nominal variables (often called 'factors'). Analysis of variance may be the most common method for the analysis of experimental data. The method exist for experiments with between-subject factors, within-subject factors, or combinations of the two. As a general recommendation, the number of factors should be kept low, preferably not more than three. The main reason is that independent variables may interact with one another, and the number of possible interactions increases rapidly when more independent variables are added to a study. Interactions are notoriously difficult to interpret, especially those that involve more than two factors. Analysis of variance is discussed in many textbooks on statistics. An exceptionally complete handbook is Winer, Brown, and Michels (1991). There are numerous examples of eye-tracking studies in which the results were analysed with an analysis of variance. One example is a study by Camblin, Gordon, and Swaab (2007), who looked at the influence of two factors on eye-movement measures. These factors were word association (whether two words are easily associated with each other or not), and discourse congruency (whether a word fits in the context or not). The main question behind this investigation was whether reading processes are more strongly influenced by local context (represented by the word association factor), or by global context (represented by the discourse congruency factor). Combining ERP measurements with eye-tracking measurements, they found discourse congruency to be a stronger factor than word association. In other words, local reading processes may be overruled by global reading processes.

Logistic regression A special case of a nominal variable is a variable that takes only two outcomes (e.g. *yes–no*, *hit–miss*, *dead–alive*). A seemingly attractive solution is to convert the outcomes to proportions or percentages. This might be allowable for the description of the data, but not for the statistical test. One risk with proportions is that some participants contribute with many data points (e.g. 90 misses out of 100 trials), whereas others contribute with only few data points (e.g. 2 out of 5). If the results from these two participants were averaged, then the first proportion would be counted just as heavily as the second, which is not appropriate since the second proportion is much less reliable than the first. The solution for such dichotomous variables is to convert the

proportional scale to a logarithmic scale (logit transformation) and to do the analysis on the transformed values instead. This type of analysis is called a logistic regression. An introduction to logistic regression can be found in Tabachnick and Fidell (2000). An example of a logistic regression analysis within eye-tracking research is given in Sporn *et al.* (2005). In that study, a number of eye-tracking variables were measured in a clinical group of schizophrenic patients and a control group. Subsequently, the results of the eye-tracking measures were used as predictors in a logistic regression analysis, to establish whether the two groups could be differentiated on the basis of the measurements.

Regression Regression is similar to analysis of variance in that there is a dependent variable measured at an interval/ratio scale. In regression, however, the factors (predictors) may be either categorical or continuous. The simplest example of regression contains one continuous dependent variable (e.g. fixation duration) and one continuous predictor (e.g. font size). A relationship between these variables implies that an increase in the predictor is associated with an increase (or a decrease) in the dependent variable. The most parsimonious representation of such a relationship is to suppose that it is linear, i.e. the change in the dependent variable is constant across the whole range of the predictor. If this is true, then the relationship between the variables can be modelled using the equation for a straight line: $Y' = b + aX$. In this equation, b is the level of Y at the lowest level of X, and a is the slope of the line, i.e. the change in Y per unit change in X. Reality may be more complex than that, however. The relationship between two variables need not be linear, and there may be more than one variable that influences the dependent variable. We recommend Cohen, Cohen, West, and Aiken (2002) as a textbook on regression.

Multilevel modelling A relatively recent development in statistical analysis is offered by so-called multilevel analysis (also known as hierarchical models, mixed models). In this type of analysis, random factors are included and parameters of the model (estimates of the contributions of the different factors) are estimated by a process of maximum likelihood estimation or variants of it. These models may be applied when the dependent variable is an interval/ratio variable, but also when the dependent variable is a nominal variable. Multilevel models have the great advantage that they are flexible. The data set does not need to be perfectly balanced, as it should be for analysis of variance. For an introduction to multilevel modelling, we refer to Singer and Willett (2003). An example of multilevel analysis within eye-tracking research is given by Barr (2008). The technique has been applied successfully to analyse results of studies with the visual world paradigm (p. 68), but its range of applications is far wider than that.

Loglinear analysis Loglinear analysis is a technique for analysing the relationship between nominal variables. If only two variables are involved, their relation can be represented as a two-dimensional contingency table. If there are three, the table becomes three-dimensional, and so on. In loglinear analysis, as in analysis of variance, the model for the expected cell frequencies consists of main effects and interaction effects. In a two-way table, for instance, there are two main effects, and one two-way interaction. In a three-dimensional table, there are three main effects, three two-way interactions, and one three-way interaction, and so on. The goal of the analysis is to find the most parsimonious model that produces expected cell frequencies that are not significantly different from the observed frequencies. An example of the application of loglinear analysis in eye-tracking research is given for transition matrices (p. 193). An introductory chapter on loglinear analysis can be found in Tabachnick and Fidell (2000).

3.3.4 Data modelling

The fourth stage, which is optional to many, is the modelling stage. In some cases, there is no noticeable difference between the analysis stage and the modelling stage. Many researchers settle for just finding individual significant effects and this is perfectly fine. However, once a particular domain has accumulated a number of significant predictors each targeting the same variable, then it becomes fruitful to try to integrate these predictors into a complete model. The aim of statistical modelling is to create an explicit model that can describe and predict your data, and do this as well as possible. This is important for the scientific work, because this output is something we can benchmark against, typically through some form of goodness-of-fit statistic. If we have two different models that try to describe a particular set of data, we can test both and see which model performs more accurately. We can also see whether the inferior model can be incorporated to create an even better unified model, or if it has no unique information value at all to contribute. The end result is a better understanding of what factors are involved in a particular behaviour and/or cognitive process, and how these factors interact to produce the outcome they do. Valuable outcomes from modelling include:

- Produce an explicit model that can be implemented in an application.
- Produce an explicit model that can be compared against other models to evaluate which one is better.
- Identify redundant factors that do not contribute with unique explanatory power.

Model-building is performed, not in a single, correct way, but rather by a variety of approaches. A typical rule of thumb is to achieve a good tradeoff between model complexity and explanatory power. Including many predictors that improve the model only minimally results in a very large and complex model. In that case it would be better to exclude those predictors and settle for a less powerful, but much simpler model. A simple model will be much easier to communicate and for other researchers to adopt.

Other questions, which really are beyond the scope of this book, are whether models should be built in a forward fashion, including predictors as they are identified as significant predictors, or in a backward fashion, excluding factors as they fail to improve the model. Different practices exist in different fields, and it is up to the reader to find her own way of modelling confidently.

3.3.5 Further statistical considerations

A potential problem that may undermine your conclusions is the *multiple comparisons problem* (see also the terms *family-wise error rate* or *experiment-wise error rate*). We briefly explained this before in the context of a fishing expedition, where we test many different measures and settle with whatever is significant. A significant result is a probabilistic statement about the likelihood that a given sample comes from the assumed population, or comes from the same population as another sample. If this probability is sufficiently low, we can reject our null hypothesis in exchange for our more interesting alternative hypothesis. However, this probability is only valid for a single test. If you test a hundred samples using this test, you most likely get a few significant tests even though the data are completely randomly generated with no real effect at all. In order for the hypothesis test to mean anything, the experiment and the analysis should be set up to make a single test for every research question, otherwise you are inflating the risk of a significant result where there is no true effect. There are several ways where a multiple comparisons problem could arise in your experiment:

- You do not have a single clear measure to capture your hypothesized effect, so you use several measures each tested with their own hypothesis test.

- You do not have a clear prediction about where in the trial an effect will appear, so you compute several time bins and test their significance separately.
- You test different layouts of the data, for example using time bins, then using the whole trial, then collapsing trials into larger units (*trial* → *block* → *participant*). There is a risk that you keep transforming and aggregating your data until your data becomes significant, rather than arranging the most appropriate way determined prior to the analysis.

One way to compensate for this problem, if indeed you want to investigate several measures, is to perform a Bonferroni correction (or related procedure, see e.g. Holm, 1979) on your significance level to compensate for the multiple comparisons. This means you lower your significance level (α) based on the number of hypothesis tests you perform. The standard Bonferroni correction is simply to calculate $\frac{\alpha}{n}$ where α is the significance level and n is the number of comparisons (hypothesis tests).

If you find effects that are only significant before the multiple-comparisons adjustment, but you still believe in them, then you can at least report them as post-hoc findings. In themselves, they are not as useful as real results, but another researcher may have a good explanation for them and proceed with her own replication of your results, if you do not do this yourself.

3.4 Auxiliary data: planning

Eye tracking is useful, fascinating, and challenging in itself, but all of these positive properties can be increased by adding further data channels. Common auxiliary data types include verbal data, reaction time data, motion tracking, galvanic skin response (GSR), and for a few years now also electroencephalography (EEG) and function magnetic resonance imaging (fMRI) data. These are added for a variety of reasons.

Verbal data, for instance, are used for methodological triangulation as an information source on cognitive processes in working memory in addition to eye tracking as information source on perceptual/attentional processes (e.g. Antes & Kristjanson, 1991; Canham & Hegarty, 2010; Charness, Reingold, Pomplun, & Stampe, 2001; Haider & Frensch, 1999; Jarodzka, Scheiter, Gerjets, & Van Gog, 2010; Lowe, 1999; Reingold, Charness, Pomplun, & Stampe, 2001; Underwood, Chapman, Brocklehurst, Underwood, & Crundall, 2003; Van Gog, Paas, & Van Merriënboer, 2005; Vogt & Magnussen, 2007. Others add verbal data to eye-movement data to study the speech processes in themselves (Tanenhaus *et al.*, 1995; Griffin, 2004; Holsanova, 2008).

All these types of data have their own possibilities, weaknesses, and pitfalls, and none of them provide an infallible turnkey solution any more than eye tracking does. Rather, the trick is to use them in combination so that the weakness of one system is complemented by the strength of the other. This is sometimes called *methodological triangulation* and *cross-validation*.

We will now describe the possibilities of using common auxiliary data in triangulation to cross-validate eye-tracking data. Eye-tracking data, including pupil diameter data can also be used to disambiguate other data, but that is outside the scope of this book.

3.4.1 Methodological triangulation of eye movement and auxiliary data

In spite of the great opportunities eye tracking provides to a researcher, it also has its shortcomings. As we noted on page 71, eye-tracking data only tell us where on the stimulus a

cognitive process operated, and possibly for how long, but not by itself *which* cognitive process is involved.

Methodological triangulation refers to the use of more than one methodological approach in investigating a research question in order to enhance confidence in the ensuing findings (Denzin, 1970). If research is founded on the use of a single research method it might suffer from limitations associated with that method or from the specific application of it. Thus, methodological triangulation offers the prospect of enhanced confidence, credibility, and persuasiveness of a research account through verifying the validity of the findings by cross-checking them with another method (Bryman, 1984). Webb, Campbell, Schwartz, and Sechrest (1966) suggested, "Once a proposition has been confirmed by two or more independent measurement processes, the uncertainty of its interpretation is greatly reduced. The most persuasive evidence comes through a triangulation of measurement processes" (p. 3). The consensus among the many reviewers of methodology, supported by empirical studies, is that it is best to rely on a wide range of complementary methods (Ericsson & Lehmann, 1996).

In psychology several methods are in use to gain data on human knowledge (for an overview see Kluwe, 1988): *probing* (i.e. interviewing a participant), *questionnaires*, *sorting tasks*, *free recall of knowledge*, as well as several behavioural measures such as *reaction times*, *electroencephalography*, *galvanic skin response*, *functional magnetic resonance imaging*, and *thinking aloud*. These additional data types vary both in how easy they are to record in combination with eye-movement data, the potential they have in disambiguating them, and in how well this potential is investigated.

Verbal data are easy to record and have a wide potential to disambiguate eye-tracking data, because this method allows researchers to gain insight into participants' experienced cognitive processes while inspecting a stimulus or performing a task. It has become the largest and most investigated complimentary data source to eye-tracking data, in particular in the applied fields of eye-tracking research, where participants have free and naturalistic stimuli and tasks.

3.4.2 Questionnaires and Likert scales

Both questionnaires and Likert scales can be seen as a form of elicitation where conscious answers are given by the participant to highly structured questions. The structure can be more or less rigid, where one extreme is open-ended questions (such as "How do you feel?"), and another extreme would be forced-choice questions with few alternatives ("Do you prefer option A or option B?"). The rigidity has both benefits and drawbacks. A great benefit is the ability to automatically have all the answers confined within an easily analysed answer space, for example values ranging between 1 and 7. A drawback of rigid questions is the risk of low validity due to wrong constructs or misinterpreted questions, such as participants not understanding what you are asking about or forced to provide an answer to a dimension they believe is irrelevant to them. Questionnaires may be low-tech, but they are critical to operationalizing difficult constructs. Assuming you want to find an eye-tracking measure that predicts the level of happiness of a participant, you will have no other easy access to such information because there exists no device that can measure the happiness of a participant. Fortunately, such questions are easily arranged in a questionnaire, especially if they are standardized questions used by psychologists. It is then easy to collect data about both the eye movements and the happiness of a number of participants, and then find a correlation with some measure which can then be further elaborated on and verified.

A typical psychological questionnaire often uses a Likert scale for easy analysability. Additionally, there are often many questions asking the same thing but with slightly different

wording in order to reduce any effect due to particular phrasing, including some reversed questions that ask the complete opposite (with a correspondingly reversed scoring). An example in line with our above example would be the Oxford Happiness Questionnaire (Hills & Argyle, 2002).

3.4.3 Reaction time measures

Eye-tracking data offer a number of reaction time measures, for instance saccadic latency, entry time, latency of the reflex blink, eye–voice latency, and others listed in Chapter 13. Even the first fixation duration is in effect a form of latency measure. All of these are indicative of processing, such that the reaction takes longer when processing is hampered or more difficult. When latency is referred to in relation to auxiliary data in this chapter, discussions are limited to identifying cognitive processes in such a 'brain sense'. Of course, there are a multitude of complex issues to do with the latencies involved with the synchronization of machines and equipment when recording auxiliary data. Chapter 4 (p. 134) and Chapter 9 (p. 286) tackle the combinations of equipment for data recording more technically. Refer back to Chapter 2 (p. 43) to remind yourself of the latency issues involved specifically with eye-trackers.

The traditional non-eye-tracking reaction time test is a measure from onset of a task until the participant presses one of two or more buttons to mark a decision, typically between two options, for instance "yes" or "no" to the question whether a series of letters constitute a word or not. The latency of the decision is then taken as the dependent variable and used as an approximation of the ease of processing of the particular stimuli. In trials where the processing leading up to the decision is easy the participant is faster, whereas hard trials have longer latencies.

As eye tracking provides the richer spectrum of latency measures, there is often little point in adding manual reaction time tests to eye tracking, other than for pure triangulation or to compare visual and manual modalities. However, time on task-data which measures the time until a participant has finished with a stimulus or subtask is often added to the analysis of eye-movement data. This additional information comes at the cost of variable trial durations, however, which requires us to think about scaling several of the other eye-tracking measures we might think about using.

3.4.4 Galvanic skin response (GSR)

Galvanic skin response (GSR) measures the electrical conductivity of the skin using electrodes which are usually put on one or two fingers of the participant. The variation in GSR signal corresponds to the autonomic nerve response as a parameter of the sweat gland function.

When eye tracking has been supplemented by GSR, the motive has been to investigate *cognitive load* and *emotional reactions*, for instance in usability tasks (Westerman, Sutherland, Robinson, Powell, & Tuck, 2007) and social anxiety research (Wieser, Pauli, Alpers, & Mühlberger, 2009).

The GSR latency is slow, reactions appear 1–2 seconds after stimulus onset. This means that the eyes could already have left the part of the stimulus that caused the GSR effect long before the effect was registered in data. This latency is difficult to take into account, and could be one reason why there are so few combined studies.

Eye tracking offers some measures of its own that are sensitive to cognitive load and emotional variations, for instance pupil dilation (p. 391) and saccadic amplitude (p. 312). Since these eye-tracking measures react to so many cognitive states (so many *c*:s in the terms of Figure 3.1 on page 72), however, disambiguating them with GSR makes good sense.

3.4.5 Motion tracking

Motion trackers can be magnetic or optic, and are used to measure the movements of all (external) body parts, but not eyes. Magnetic motion trackers are sometimes optional parts of head-mounted eye-trackers. Optical motion tracking is based on infrared cameras and reflections just like eye tracking, and gives the same type of sample data stream, with comparable sampling frequency and precision, albeit 3D, for a selected number of points across the participant's body or on artefacts manipulated by the participant. Even the analysis of general movement data has similarities to fixation and saccade analysis. The obvious benefit of adding motion tracking to your study is that you will be able to measure synchronized movements of the eye, body, and objects.

The combination is not uncommon in applied research. For instance, Wurtz, Müri, and Wiesendanger (2009) investigated the eye–hand latency in violin players, as the interval from the fixation of a note until the corresponding bow reversal, and Wilmut, Wann, and Brown (2006) investigate the role of visual information for hand movements. Wengelin *et al.* (2009) and Andersson *et al.* (2006) describe set-ups for studying how reading of one's own emerging text coincides with keyboard writing (keylogging), and Alamargot, Chesnet, Dansac, and Ros (2006) have developed a set-up and software solution called Eye and Pen, which combines graphomotor activities with eye movements. There are also many human factors, ergonomic, and robotic applications of this combination.

3.4.6 Electroencephalography (EEG)

There are many similarities between electroencephalography (EEG) and eye tracking: sampling frequencies are in the same range, and both signals can be analysed as process measures. There are different EEG technologies (called high- and low-impedance) that require different post-processing. And with both measurement techniques, it takes some time to gather enough experience to be able to do publishable research.

EEG does not measure deep into the brain, only the surface, and there is high inter-individual variance in the thickness of the skull and scalp. High amplification is needed as the signals are often very weak. Because the noise levels are so high, many trials are needed to filter out a significant effect, and participants may find this tedious. EEG artefacts stem from alternating current but also eye blinks and saccadic and microsaccadic movements. Filters are required to remove them, and high-impedance systems may require heavier filtering.

Sampled EEG data come in waves that correspond to continuous brain activity. It is possible—with some training—to read state of arousal directly from wave plots, which is extensively done in clinical settings (hospitals). When we are excited and alert, the signal is high in frequency (Hz) and low in amplitude (μV). When we are drowsy, the activity is much slower but higher in amplitude. EEG can be analysed in the frequency domain in order to extract information about the global brain activity of the participant.

When EEG is added to eye tracking, the continuous EEG signal is seldom used. Instead analysis focuses on the EEG amplitude, direction (positive/negative), and latency of the signal with a particular scalp distribution *as a response* to external stimulus events or internal cognitive processing. This is called *event-related potentials* or ERP (Luck, 2005).

In one line of research, the purpose has been to study the neurological system itself. The *saccadic eye-movement-related potentials* (SERP) and the *eye-fixation-related potential* (EFRP) are ERP paradigms that investigate the EEG signal next to saccades and fixations. Early studies focused on the neural activity around saccades, and what that could tell us about the human visual system. Becker, Hoehne, Iwase, and Kornhuber (1972) found that around 1–3 seconds before the saccade onset, occipital and parietal posterior areas exhibit a so-called pre-motion negativity, indicative of general readiness. A pre-motor positivity can

be measured 100–150 ms before the saccade onset, possibly reflecting motor programming (Jagla, Zikmund, & Kundrát, 1994). Immediately after the saccade offset, there is a strong positive response, called the Lambda response, in the posterior parietal area, and a concurrent negativity in the frontal eye fields. The shape of the Lambda response depends on the general visual background (Morton & Cobb, 1973), and is believed to correlate to the processing of new information in the visual cortex. The negativity in the frontal eye fields is probably a sign of inhibition of further saccades while the processing of information continues in the visual cortex (Jagla *et al.*, 1994). For a recent review of research on *saccadic eye-movement-related potentials*, see Jagla, Jergelová, and Riecanský(2007).

In reading research, ERP data are used to support and strengthen interpretations made from eye-tracking data. Dambacher and Kliegl (2007) found a correlation between N400 components and fixation durations. Takeda, Sugai, and Yagi (2001) found that the EFRP in the 100–200 ms block after fixation onset decreases in a way that would reflect decline of mental concentration (i.e. carelessness) caused by visual fatigue. Using the same P200 EFRP, Simola, Holmqvist, and Lindgren (2009) show a parafoveal preview benefit for distinguishing between words and non-words in the right visual field that does not exist in the left visual field.

3.4.7 Functional magnetic resonance imaging (fMRI)

Functional magnetic resonance imaging (fMRI) measures activity throughout the whole brain, not just surface activity like EEG. The *temporal resolution* differs very much between fMRI and eye tracking. Eye tracking involves measuring how the eyes move with a temporal resolution of down to 0.5–1 ms. In contrast, fMRI involves measuring and aggregating over 1000 ms time spans. This makes it more difficult to co-analyse fMRI and ET data than EEG and ET data, where both systems have the same temporal resolution. The output from an fMRI measurement is an activation visualization of the blood oxygenation level-dependent (BOLD) signal, which in principle is identical to a heat map and the eye movement representations of Chapter 7.

Although fMRI studies very often include looking at pictorial stimuli (and/or hearing audio), the vast majority of studies that combine the two technologies only use eye tracking to control that the participant is awake, has his eyes open and looks in the general direction of the stimulus. If researchers analyse the eye-tracking data, they usually only detect saccades, and only to make sure that the eye is not moving.

As an example, Simola, Stenbacka, and Vanni (2009) measured activity in the visual cortex (V1) as participants *looked at* a central cross on the stimulus monitor and simultaneously *attended to* wedges in five concentric rings at 1.6°–10.2° from the centre. The authors showed that the enhanced activity by attention in retinotopically organized V1 directly corresponds to the locus of covert attention, and that the attended responses spread over a signicantly larger area than the sensory responses. It was important to show that the participant's gaze did not deviate systematically from the central cross, because eye movements would move the mapping of the stimulus onto the visual cortex (V1). Hence eye tracking was used.

A rare exception where saccades were actually used to align the fMRI data is Ford, Goltz, Brown, and Everling (2005). They used an antisaccade task with long intervals between saccades, which is compatible with the slow fMRI data.

3.4.8 Verbal data

This section describes the most commonly used method for knowledge elicitation in combination with eye tracking: verbal data. We use the term *verbalization* for the act of external-

izing thoughts as speech and *verbal data* for the totality of data resulting from recordings of verbalization, irrespective of their form (i.e. audio or transcribed). Combined recordings of eye tracking and verbal data are made in several research areas as well as in applied usability projects. There are three major purposes to record eye-tracking data in combination with verbal data:

1. To investigate the minute relation between vision and speech over time (Holsanova, 2008).
2. For purposes of methodological triangulation, for instance to investigate working memory processes directly in addition to perceptual/attentional processes as shown by eye-tracking data (Jarodzka, Scheiter, *et al.*, 2010; Altmann & Kamide, 2007).
3. In specific cases, eye-tracking data are recorded to help participants to elicit verbal data by a method known as "cued retrospective reporting" (Hansen, 1991; Van Gog, Paas, Van Merriënboer, & Witte, 2005).

Theoretical background: origin and idea of verbalizations as a valid data source

Initially, the easiest and most common way to gather insight into cognitive processes accompanying task performance was to interview people who are skilled performers, that is experts (Ericsson, 2006). It is questionable, however, whether experts are able to describe their thoughts, behaviours, and strategies so that it is understandable to less skilled people (Ericsson, 2006). In particular, since discrepancies have been found between reported and observed behaviour (Watson, 1913). For this reason Watson (1920) and Duncker (1945) introduced a new method of thought analysis: *thinking aloud*. This type of verbalization has been shown not to change the underlying structure of the thoughts or cognitive processes, and thus avoids the problem of reactivity, as long as the verbalizations are carefully elicited and analysed (Ericsson & Simon, 1980, 1993).

The central assumption behind the use of thinking-aloud is that "it is possible to instruct participants to verbalize their thoughts in a manner that does not alter the sequence and content of thoughts mediating the completion of a task and therefore should reflect immediately available information during thinking." (Ericsson, 2006). Those verbalizations provide data on which knowledge is currently activated and how it changes. According to Ericsson and Simon (1993) the information processing model assumes the following: (1) the verbalizable cognitions can be described as states that correspond to the contents of working memory (that is, to the information that is in the focus of attention); (2) the information vocalized is a verbal encoding of the information in the working memory. That is, only this content can be found in the data that was "on the participant's mind", respectively in the participant's attention. It is important to note that if thinking aloud is not completely free, it may interfere with task performance itself. Providing the participant with appropriate instructions is therefore crucial (p. 105).

Another crucial part in the use of verbal reports is coding (Chi, 2006 and page 290). The data should be coded in the context of the task. Hence, a cognitive task analysis needs to be done beforehand, so as to know the functional problem states required to be able to categorize single utterances.

Thinking aloud techniques have been successfully used in a variety of domains, like designing surveys (Sudman, Bradbrun, & Schwarz, 1996), learning second-language (Green, 1998), text comprehension (Ericsson, 1988; Pressley & Afflerbach, 1995), decision-making studies (Reisen, Hoffrage, & Mast, 2008), studies of text translators (O'Brien, 2006), developing computer software (Henderson, Smith, Podd, & Varela-Alvarez, 1995; Hughes & Parkes, 2003), or to investigate the relation between vision and speech (Holsanova, 2008).

This method can provide information on, for instance, the forward-strategy-use in experts (Smith & Good, 1984), or even in perceptual processes; for example it has been found that experts note more relevant features of pictures in contrast to novices (Wineburg, 1991).

It has to be noted that the method of gathering verbal data from participants is known under several other names, such as a *retrospective think-aloud* in the academic usability world (Hansen, 1991; Hyrskykari *et al.*, 2008) and as a *post-experience eye-tracked protocol* (PEEP) in the commercial usability world (Petrie & Harrison, 2009; Ehmke & Wilson, 2007). We decompose the term verbal reports (Ericsson & Simon, 1993) into:

- Thinking aloud approaches (like concurrent reporting, retrospective reporting, and cued retrospective reporting (Van Gog, Paas, Van Merriënboer, & Witte, 2005).
- Probed reporting, like self-explanations (i.e. the participant explains a stimulus or task to himself; Renkl, 1997)
- Structured interviews.
- Free recall.
- Task-driven verbalizations (i.e. providing verbalizations according to a specific task).

Individual differences in verbal data

Already Claparède (1934) and De Groot (1946/1978) had found large differences among participants in their ability to think aloud. To give you an impression of what variation in participant verbosity can be expected, we present here frequency distributions from real data. Figure 3.8 presents data from a study, where we used cued retrospective thinking aloud (Jarodzka, Scheiter, *et al.*, 2010). Thus, participants are very likely to vary in how much verbalization they produce. Although this situation cannot be completely avoided, it helps to train the thinking aloud and to prompt silent participants when they stop talking (see below).

Forms of verbal data

In this section, we will distinguish between different forms of verbal data according to the point in time when they have been produced: concurrently or retrospectively. Thinking aloud, self-explaining, and task-driven verbalizing are produced *during* stimulus inspection (concurrent verbalizations). Retrospective reports (i.e. reflecting what the person was thinking during stimulus inspection) can also be produced *after* stimulus inspection as well as free recall and structured interviews (retrospective verbalizations). Table 3.2 provides an overview of their properties.

Concurrent verbalizations

Thinking aloud can be produced in two points in time: concurrently or retrospectively. If the participant speaks while performing a task or inspecting a stimulus, this set-up is called *concurrent think-aloud*. Meanwhile his eye movements can be recorded. Most eye-tracker manufacturers have support for synchronized concurrent recordings of speech, but the very act of speaking may make the participant quiver or move enough that the recording of eye movements will be less precise, in particular for tower-mounted eye-trackers (p. 137). Moreover, it is very likely that participants thinking aloud perform slower (Karpf, 1973).

It has long been suspected that concurrent verbalizations alter eye movements during the task. On the one hand, psycholinguistic research in the so-called "visual world paradigm", starting in the mid 1990s (Tanenhaus *et al.*, 1995) has thoroughly investigated the temporal relation of gaze to verbal expressions. Its main thesis, that "speech is timelocked to gaze" has been shown for single-sentence trials, again and again. However, in the task of describing complex pictures with everyday scenes, speech planning is a process in itself, which in turn requires additional time and affects eye-movement behaviour (Holsanova, 2001, 2008).

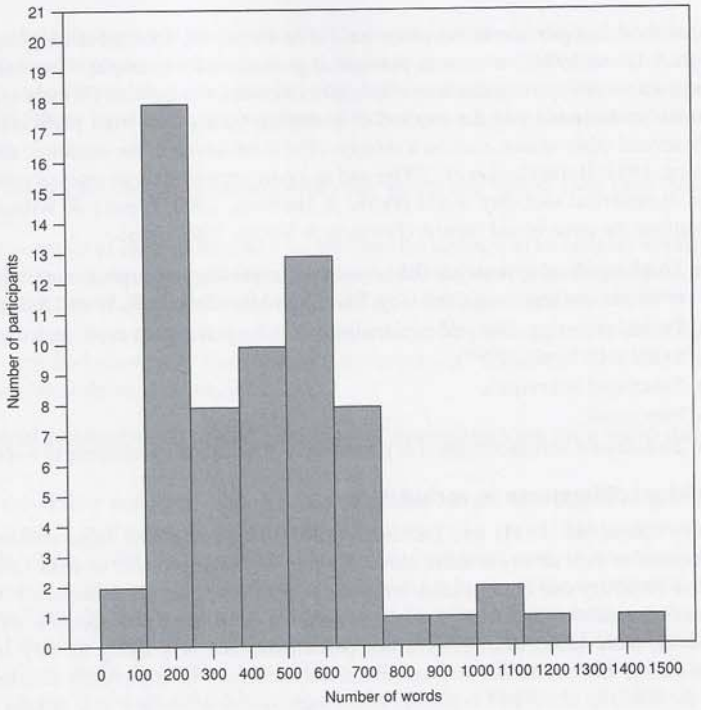


Fig. 3.8 Number of words that participants uttered during cued retrospective thinking aloud. Bin size is 100 words. Verbalization training and prompting was used in the way later described in this section. Data from Jarodzka, Scheiter, *et al.* (2010).

Ericsson and Simon (1993) claim that, given that thinking aloud is implemented in the manner they propose, thinking aloud should not alter task performance itself, besides slowing it down. Nevertheless, some researchers found exactly this effect: the think-aloud process takes resources from all parts of the cognitive system, and slows down not only eye movements, but the general exploration and learning processes (Nielsen, Clemmensen, & Yssing, 2002; Van Someren, Barnard, & Sandberg, 1994). Eger, Ball, Stevens, and Dodd (2007) found that fewer participants finished their online search task when thinking aloud compared to being undisturbed during the task. Davies (1995) even found that the order in which the participant performs subprocesses changes when think-aloud is required of him in a design task. The greater the cognitive load a task imposes, the more novices have problems with concurrently thinking aloud compared to experts (Van Gog, 2006).

The advantages, on the other hand, are the following. Two data sources may be recorded at one time. These data sources are very likely to be closely linked, since they have been recorded simultaneously from a single participant. Concurrent verbalization also provides the momentous perspective. This would be of particular importance in complex tasks, where cued retrospection could be expected to provide a perspective that deviates from or even ignores the momentous cognitive processes and simply becomes a post-hoc construction. This happened for Ryan and Haslegrave (2007) who showed videos (without gaze data) of workers in a storage room, and collected retrospectives.

Concurrent verbalization is used frequently in psycholinguistics research, where the very

Table 3.2 Overview of the different varieties of verbal reports that are combined with eye-movement data, and their properties as methods. Note: Y – yes or possible, N – no or very unlikely.

Method	Dual recording sessions necessary?	Verbal and eye-movement data synchronized?	Pre-structuring of verbal data possible?	Decreased eye-tracking data quality?	Task performance slower?	Effect on task performance?	Risk of memory loss on verbal data?	Biased verbal data?
Concurrent recording								
concurrent thinking aloud	N	Y	N	Y	Y	Y?	N	N
self-explanation	N	Y	N	Y	Y	Y	N	N
task-driven speech by describing stimulus	N	Y	N	Y	Y	Y	N	N
Retrospective recording								
retrospective reporting	Y	N	N	N	N	N	Y	Y
cued retrospective reporting	Y	N	N	N	N	N	Y	N
structured interviews	Y	N	Y	N	N	N	Y	Y
freely recalling the content of a stimulus	Y	N	N	N	N	N	Y	N

purpose is a detailed investigation of the temporal relation of visual attention to the contents of verbal data, and an investigation of the implications for speech production and reception models. Furthermore, this method is frequently used in educational psychology to investigate levels of processing involved in studying certain learning materials as well as mental overload. The suspicion that the primary task may be affected has discouraged many other applied researchers from using concurrent verbalization, however.

Besides thinking aloud, at least two more types of verbal reports exist that may be linked to eye movements: *self-explanations* and *task-driven verbalizations* (in psycholinguistics). Self-explanations are a specific variety of verbalizations which require the participant to explain the stimulus to himself. Whereas thinking aloud should not interfere with the primary task, this type of verbalization is meant to alter the task performance or the inspection of the stimulus. This method is mainly used in educational psychology, where it has been shown to enhance learning (Chi, De Leeuw, Chiu, & LaVancher, 1994; McNamara, 2004; Renkl, 1997). Since this kind of verbal data is recorded concurrently, it possesses similar advantages and drawbacks to the concurrent thinking-aloud method.

Task-driven verbalizations are another specific form of concurrent verbalizations. This method requires that the participant does not freely think aloud, but has a specific task in mind, like describing or recalling a stimulus. Again, this type of verbalization will change the eye-movement performance dramatically as compared to a silent stimulus inspection.

Retrospective verbalizations

The alternative to concurrent recordings is to record the thinking aloud *after* the task is performed. Separating the eye-movement recording during the primary task in time from the verbal recordings could make the study liable to loss of detail from memory as well as fabulation. The question of whether participants remember or fabulate when they explain their own eye movements was elegantly answered by Hansen (1991) who showed participants bogus recordings of someone else's eye movements. The misled participants soon detected the error, which Hansen took to indicate that they could remember their own eye movements. Participants' intact memory is further supported by Guan, Lee, Cuddihy, and Ramey (2006), who find that participants look at objects in the same order as they later (even without support from eye movement data) say that they do. Moreover, several studies have shown that *retrospective think-aloud* results in more detailed and qualitatively better verbalizations if combined with showing the participant's own eye-movement recordings compared to uncued retrospective verbalizations (Hansen, 1991; Van Gog, Paas, Van Merriënboer, & Witte, 2005). The verbalizations of the cued verbalizations were quantitatively better in terms of eliciting more information on actions done, more descriptions of how a step was performed (Van Gog, Paas, Van Merriënboer, & Witte, 2005).

This method exists in several varieties that go under names such as *cued retrospective reports* (Eger *et al.*, 2007; Hansen, 1991; Van Gog, Paas, Van Merriënboer, & Witte, 2005), *eye-movement supported verbal retrospection* (Hansen, 1991), or *post-experience eye-tracked protocol* (Petrie & Harrison, 2009; Ball, Eger, Stevens, & Dodd, 2006), reflecting the fact that the method has been re-discovered more than once in different fields of research.

An important issue to consider is that a whole body of studies showed that cued retrospective verbalizations stimulate meta-cognitive reflection at the cost of action-related comments. Hyrskykari *et al.* (2008), in a test of web usability, found that cued retrospection resulted in more comments on the user's cognitive processes, while think-aloud results in more comments on user manipulation (of the software/web pages). Eger *et al.* (2007) found that more usability problems were identified by participants who performed cued retrospection compared to think-aloud or playback of screen without gaze data. Taylor and Dionne (2000); Kuusela and Paul (2000) found that more action and outcome statements are produced in concurrent think-aloud then in retrospective mode, which gives information about strategies and reasons for actions. Hansen (1991), in an analysis of computer interfaces, found that verbal retrospective protocols cued by an eye-movement video of the user's work are superior to retrospective protocols primed by a pure video recording. Hansen found more problem-oriented comments and more comments on manipulation in the task, when recording cued retrospection from participants that see their own eye movements compared to seeing just a video recording. Kuusela and Paul (2000) also argue that retrospective reports often only reveal those actions that led to a solution, and that attempts that led nowhere are not mentioned. Van Gog, Paas, and Van Merriënboer (2005) found that cued retrospectives result in a larger number of metacognitive comments (on knowledge, actions, and strategies of the participant), and that they elicit more information on actions done, more descriptions of how a step was performed (Van Gog, Paas, Van Merriënboer, & Witte, 2005).

The drawbacks of the retrospective method are that recordings take at least twice as long as with concurrent reporting and that the two data sources may not be as perfectly synchronized as in concurrent reporting. Moreover, if the task is too long, participants may easily forget what they have been thinking even when cued with their own eye movements. As a rule of thumb reported by researchers using this technique, recordings should not exceed ten minutes (Van Meeuwen, 2008). On the other hand, the main task performance is not disturbed by a secondary task (thinking aloud), which in turn may result in a more naturalistic

task performance or stimulus inspection.

Another version of verbal data is *free recall* (e.g. Jahnke, 1965). In this, a participant simply recalls the stimulus in a free order and without cues. Free recall is used with eye tracking as an experimental condition in mental imagery studies (Johansson *et al*., 2006).

Another possibility of recording verbalizations is that the researcher prepares questions that she uses in *a structured interview based on gaze replays* with the participant (e.g. Pernice & Nielsen, 2009; Ehmke & Wilson, 2007). Questions should be designed as part of the experimental design and the interview should have the same structure for participants. Sometimes, however, the participant and the researcher look through the scanpath or gaze cursor playback together, simply discussing whatever strikes them as interesting, with or without prepared topics.

In usability, the joint discussion or interview is often done after the researcher previews the participant's data, before letting the participant see it, as a means of coding the scanpaths so the right questions can be asked in terms of the cognitive processes underlying the data (Pernice & Nielsen, 2009; Ehmke & Wilson, 2007). Before using this method, note that you may easily run the risk of including three severe drawbacks to your data. First, the coding of the scanpath plot is subjective. No algorithms exist as yet that would detect such patterns in real time. This means that different measurements of the same scanpath would not lead to the same questions. Thus, this measure is not *reliable*. Second, the time the participant has to wait until he can be questioned about his proceeding may easily be too long to deliver a trustworthy recall of the process. Since working memory is very limited in time, a long pause between action and recall (without memorizing) leads to forgetting the content. Since, participants are not asked to memorize their thoughts, they will not be transferred to long-term memory. This means that the measured verbalizations are not about the intended content, instead they are very likely to be made up. Under such circumstances this measure would be not *valid*. Third, if the coding of the scanpath is conducted by only one rater under time pressure depending on which experimenter is conducting the study on a certain day, the results may differ. Therefore, instructions for the coding have to be very strict and avoid subjectivity, otherwise, the measure is not *objective*. Hence, this method violates all three quality factors that a measurement must have (e.g. Lienert & Raatz, 1998).

Importance of instruction: how to elicit verbalizations from participants

Technical issues on how to record verbal data have been described in Chapter 4. Here we focus on the main challenge in providing valid verbal data, namely on how to elicit valid verbalizations from participants by the appropriate instruction. This issue is most crucial for two forms of verbalization: thinking aloud and self-explanations.

When recording verbal data and eye movements to retrieve current or remembered cognitive processes from the participant, the precise instruction to verbalize is of great importance. The instruction is done in three steps: *instruction*, *training*, and *reminding*. The three steps differ slightly depending on whether you record thinking aloud or self-explanations.

What both types of recordings have in common is that very sensitive data is recorded, namely speech without the ability to modify anything. Many people feel uncomfortable if their own voice is recorded. Even more important, telling ones own thoughts is quite intimate and requires a degree of meta-cognitive awareness and self-confidence. Thus, it is important that the participant feels secure and comfortable during recording. The training before a first recording helps to get them familiar with the situation. Moreover, it helps when as few people as possible are in the recording room, so the participant does not feel monitored.

It is important to emphasize in the instructions to think aloud to express thoughts freely and not with any specific task in mind (like evaluating the stimulus). Only such instructions minimize the effects on task performance. Note that the instructions to think aloud are gen-

eral instructions, thus they are suitable for both concurrent and retrospective reporting (e.g. Ericsson & Simon, 1993).

Instruction

The instruction to *think aloud* is very important, since it tells the participant what to do. Thereby, the emphasis should be on expressing the content of working memory with as little filtering as possible. That is, effort in formulating grammatically correct sentences or meaningful content should be forgone. Only such instructions can assure that the participant is not too disturbed in his primary task performance. The following instruction has been proven to elicit the desired behaviour (Van Gog, Paas, Van Merriënboer, & Witte, 2005, based on Ericsson & Simon, 1993):

> Thinking aloud means that you should really think aloud, that is, verbalize everything that comes to mind, and not mind my presence in doing so, even when curse words come to mind for example, these should also be verbalized. Act as if you were alone, with no one listening, and just keep talking.

The instruction to *self-explain* could be as follows (adapted from Van Gog, Paas, Van Merriënboer, & Witte, 2005).

> Research has shown that learning is more effective when you self-explain the to-be-learnt content to yourself. Verbalize your self-explanations always out loud as you would talk to yourself and do not mind my presence in doing so. It is not important that the self-explanation is well formulated, even when curse words come to mind for example, these should also be verbalized. Act as if you were alone, with no one listening, and just keep talking.

Training

The training, i.e. getting acquainted with *thinking aloud* gives the participant an impression of what is meant by thinking aloud and enables him to get used to the recording situation. There are at least two common training tasks that are suitable to train thinking aloud (Ericsson & Simon, 1993):

> Please think back to the home you were living in when you were a child, and count the number of windows it had while thinking aloud, verbalising everything that comes to mind.

> Please, multiply the numbers 23 by 16 and tell me, what you are thinking during your calculation.

For both tasks, the participant should count out loud stepwise (instead of giving the result immediately) and think out loud all the while. If a participant does not manage to think aloud, the other task should be tried.

The training is a central part of *self-explanation*. Research has shown that only a successful self-explainer profits from this verbalization method. A direct intervention to foster self-explanation is to train the verbalization itself (for indirect methods see e.g. Catrambone, 1998; Renkl & Atkinson, 2003; Renkl, Stark, Gruber, & Mandl, 1998). Although, several extensive training types exist (e.g. McNamara, O'Reilly, Rowe, Boonthum, & Levinstein, 2007), we present here only a very simple version of self-explanation training (Renkl *et al.*, 1998):

- Before the actual recording the experimenter models a self-explanation behaviour on a task that is comparable to the experimental task. The experimenter has to give hints on how to self-explain the given problem and to elicit several aspects of self-explanation

(e.g. elaborating the problem given, principle-based explanations, goal-operator combinations).

- On the basis of this warm-up, hints to self-explain the rationale of the presented solution steps have to be given. The hints should focus on the subgoal of each step and the operator used to achieve it (i.e. explanation of goal–operator combinations).
- Afterwards, the participant has to self-explain on his own on another comparable task, whereby he is coached by the experimenter. The coaching procedure consists of two elements:
  - If important self-explanations are omitted, this is indicated and the participant is asked to supplement the missing explanations;
  - the experimenter answers the participant's questions concerning the self-explanations.

Prompting

Both when answering questions and when narrating freely, participants will vary in how much speech they produce (p. 102). If the participant stops verbalizing his thoughts, the experimenter has to *remind* him after 3 seconds (Van Gog, Paas, Van Merriënboer, & Witte, 2005); other researchers use even longer time spans of 15 seconds (Renkl *et al.*, 1998) by saying: "Please try to keep talking."

The difference in time until prompting reflects what you want to elicit from the participants. If you are intereseted in working memory content, each silent second is missing data. Hence, you should prompt the participant to talk as soon as possible.

Prompting a participant may interfere with the task, in particular with eye movements, even when done neutrally (Kirk & Ashcraft, 2001), and should therefore be used very carefully in *concurrent verbalization* mode. Ericsson and Simon (1993) recommend the use of non-directive prompts such as "keep talking" if the participants fall silent, but not to intervene in any other way. In practice, many usability practitioners instead use interview-like prompts such as "what do you think it means?", which is likely to disrupt the flow of task processing and change eye movement and other behaviour (Boren & Ramey, 2000).

During *retrospective* thinking aloud, we have two different cases: if the gaze path is shown as a dynamic eye-movement visualization, then time is running on the monitor used for cueing just as much as it was during the original performance of the task. Prompting in such a case may interrupt the retrospection just as much as it interrupts concurrent thinking aloud. If the eye movements are shown as a static visualization, then there is no time running. Prompting in this case can very well be made. In the case where the participant stops verbalizing his thoughts, the experimenter has to *remind* him after three seconds by saying the same as above: "please try to keep talking." (Ericsson & Simon, 1993).

Do I really have to stick to those stiff instructions?

Sometimes studies recording verbal data reveal contradictory results. One important reason for that might be the actual use of an instruction to think aloud. In the usability world it is common to use more directed instructions to elicit verbal data. That is, participants do not simply mention what comes into their minds, but rather they are instructed to "evaluate" a material. That kind of verbalization, however, requires a lot of cognitive resources from the participant and thus is very likely to disturb the primary task and cause change to the content of the verbalizations (compare level-3-verbalizations; Ericsson & Simon, 1993).

A study by Gerjets, Kammerer, and Werner (2011) investigated this issue directly. The authors compared the verbal data of participants who either received an instruction to verbalize freely according to Ericsson and Simon (1993) or an explicit instruction to mention factors that influence their evaluation as often used in web research (e.g. Crystal & Greenberg, 2006;

Rieh, 2002; Savolainen & Kari, 2005; Tombros, Ruthven, & Jose, 2005. Results show that both groups differed significantly from each other in terms of verbal reports, eye-tracking data, and problem solving data. Obviously the natural behaviour was altered.

In the most applied eye-tracking fields, however, some practitioners do not place a high emphasis on the instruction to the participants. For instance, Pernice and Nielsen (2009, pp. 113–114) show examples of participants' awe-struck comments on the gaze cursor ("I can't believe that's my eye"). This may be an effect of poor instructions. The authors then quote a usability analyst who argues in favour of previewing the data to decide questions that participants can be asked, apparently oblivious of the danger of fabrication and biases. It is not uncommon that applied users fail to apply a sufficient methodological standard to their use of the retrospective method, and later mistakenly attribute their failures to the method rather than to their own standards. As with all scientific methods, retrospective verbal protocols also require methodological rigour.

Thus, different instructions to verbalize thoughts lead to differences in verbalizations, eye movements, behaviour, and level of disturbance of the primary task. Since, the very free instruction of Ericsson and Simon is the most examined and elaborated one, its consequences can be estimated. Whereas, if you make up your own instruction, you never know, what comes with it. Thus, we recommend to use the free instruction, in particular, since it does not disturb the primary task.

3.5 Summary

This chapter has introduced the most important parts before you proceed to record actual data. There are good reasons for spending time at the design phase of your experiment:

- Selecting how you will **approach** your research in this experiment determines the work you need to do, whether this is by an exploratory pilot, a fishing trip, a theory-driven experiment, or a paradigm-bound experiment. These approaches all have strengths and weaknesses.
- Mapping out the **logic** behind the experiment saves you the moment of despair when you realize that your study was built on false premises or a fallacious argument just as you were getting ready to write up your results.
- Selecting the correct **measures** is a decision best taken at the design stage. There has to be a clear motivation for a measure, with a theory or at least a plausible explanation linking the eye movement to the cognitive process being studied.
- The **statistics** should be prepared and tested before you record the actual data. In many cases, it is simply easier to design around a particular statistical method rather than having to learn and implement some advanced statistical analysis to cope with non-standard data.
- If required, there is always the option to **triangulate** your construct of interest using other data sources that complement the eye tracking. However, the price of this is increased complexity to your experiment.

The experimental design is perhaps the most important stage of all, and it is difficult to sum up briefly. It is all too easy to jump right in and start recording, thinking you can sort the rest later. With experience, and a couple of poor experiments, however, the lesson is learnt. A week spent properly thinking about the design saves four weeks of frustration during the analysis stage and when writing up the paper.

There are many pitfalls. For example, do the data have a distribution that is easy to work with and compatible with the statistical tests in mind? Is your selected eye-tracking measure

actually measuring what you are interested in, or are there better candidates? When in doubt, record some pilot data and take it from there.