Table 2.1 Properties of the recording environment and skills of those who run it.

Property to check	Risks if you ignore the property
Cramped recording space	Uncomfortable participant
Lab availability	Difficult to get participants to show up
Sunlight, lamps and lighting conditions	Optic artefacts, imprecision, and data loss
Electromagnetic fields	Optic artefacts, inaccuracy, and data loss (magnetic headtracking systems)
Vibrations	Variable noise, low precision
Scientific competence of technical staff	Invalid, unpublishable results; time-consuming studies
Recording experience of staff	Data quality low
Programming experience of staff	Data analysis very time-consuming
Statistical experience of staff	Invalid, unpublishable results; confusion

2.4 How to set up an eye-tracking laboratory

An eye-tracking laboratory needs both physical space for the eye-tracker and the experiments, and an infrastructure that keeps the laboratory up to date and running.

2.4.1 Eye-tracking labs as physical spaces

There is not one single solution for designing an eye-tracking laboratory. Every place where there are active people can be made into a place where researchers eye-track people. Take a car with a built-in eye-tracker and other measurement systems, or the mobile eye-trackers that we used in supermarkets for a study of consumer decision making. Neither are labs in the traditional sense. So, what is an eye-tracking lab, and how should it be designed? Most researchers work with single monitor stimuli, rather than real-life scenes. They then, in the authors' experience, prefer sound and light isolated rooms, minimizing the risk of distracting participants' attention from the task. They also tend to put their eye-tracker in very cramped locations (cubicles), where there is little room to turn around, let alone rebuild the recording environment for the needs of different studies.

In our lab, we found it useful to make the windowless recording rooms large enough (around 20–25 m^2) to be able to rebuild their interior depending on the varying needs of different projects (see Figure 2.1). Many labs—including our own—have also built one-way mirror windows between recording rooms and a central control room. This allows the researcher controlling the experiment to leave the participant(s) alone with their task, whilst still being able to monitor both recording status on the eye-tracking computer, and the participant through the one-way mirror. Having several recording rooms allows for multiple simultaneous recordings. At our lab in Lund, this has proved valuable more than once, when large data collections are to be made in a short period of time.

It is useful to minimize direct and ambient sunlight (i.e. to have few or no windows), and to illuminate the room with fluorescent lighting (the best are neon lights), which both emits less infrared light and vibrates less than incandescent bulbs (the worst are halogen lamps). Figure 4.15(a) on page 126 shows what a halogen lamp can do—note, do not make the room too dark, as this makes the pupil large (and variable), affecting data quality for most eye-trackers. A bright room keeps the pupil small even with a variable-luminance stimulus, which generally makes the data quality better. Also, in darker rooms the participant may

Table 2.2 Eye-tracker properties to ask manufacturers about.

Property to check	Risks if you ignore the property	Page
Manufacturer staff and openness policy	Poor support; strange errors in the system that are not explained to you	15
Manufacturer major user groups (publications; visits)	System properties that you need may be lacking	12
Software upgrade cycles and method	No improvement in software for years; a lot of hassle with software details	-
What eye movements can the system measure?	Study impossible to operationalize	23
Bi- or monocular	*Small* differences in fixation data go unnoticed	24
Averaging binocularity	Large offsets when one eye is lost	60
The quality of the eye camera	Noise (low precision)	38
Can the eye image be seen?	More difficult to record some participants; poorer understanding of system	116
The gaze estimation algorithm	Low data quality (precision and accuracy)	27
Frequency of infrared used	Poor data outdoors and in total darkness	-
Sampling frequency	You *may* need to record much more data; velocity and acceleration values invalid	29
Accuracy	Spatial (area of interest) analyses will be invalid	41
Drift (accuracy drops over time)	Constant recalibration; experimental design changes	42
Precision	Fixation and saccade data will be invalid; gaze contingency difficult; small movements not detectable	34
Filters used in velocity calculations	Fixation and saccade data will be imprecise	47
Headbox (remotes)	Data and quality loss when participant moves	58
Head movement compensation algorithm	Noise (low precision); spatial inaccuracy	-
Recovery time	Larger data loss just after participant moves or blinks	53
Latencies (in both recording and stimulus software)	Invalid results; gaze contingency studies impossible	43
Camera and illumination set-up	Data recording difficult or not possible with glasses	53
Robustness, the versatility for recording on more difficult participant populations	Data loss and poorer data for many participants	57
Portability of mobile system	Cannot be used out of laboratory	-
Connectivity	Difficult or impossible to add auxiliary stimulus presentations or data recordings	-
Tracking range	Data loss when participant looks in corners	58
Reference system for output coordinates	Data analysis very time consuming for some head-mounted systems	61
Parallax	Small and systematic offset in gaze-overlaid video data	60

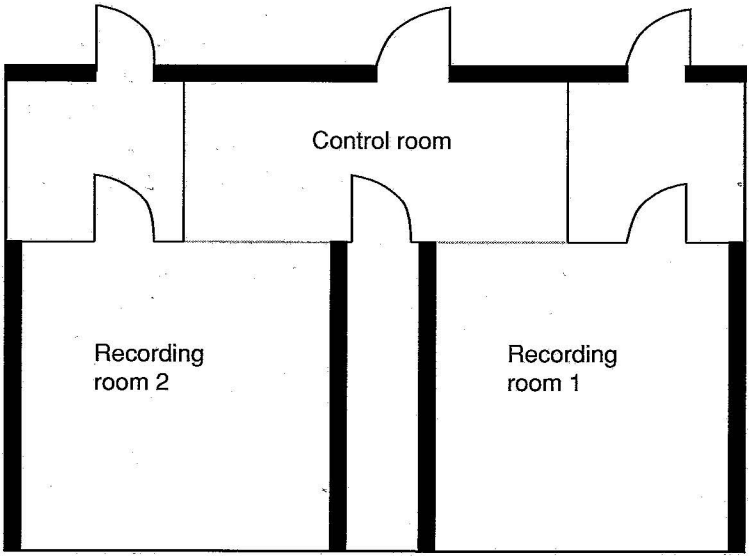


Fig. 2.1 Layout of one recording area in the Humanities Laboratory at Lund University. Each recording studio is 25 m^2. Ante-chambers allow for reception of participants, storage, and a space for the researcher to work between participants. We also have several additional recording rooms for minor studies, technical maintenance, and storage.

see the infrared illumination reflected in the mirror, although this depends somewhat on the wavelength of light emitted from the illumination.

Sounds will easily distract your participant's visual behaviour, so it is advisable to use a soundproof room if you can. For sensitive measurements, place the eye-tracker on a firm table standing on a concrete floor. Do not allow the participant to click the mouse or type on the keyboard on the same table where the eye-tracker is located. Also minimize vibrations from nearby motion of people or outside traffic. If you are using magnetic systems for head-tracking, also minimize various sources of electromagnetic noise (lifts, fans, some computers) in the recording and neighbouring rooms. Stabilized electrical current is an advantage for some measurements, but not critical.

If you are recording eye-tracking data in fMRI systems, the strong magnetic fields require optimized eye-tracking equipment to be used, typically built to film the eye from a safe distance with long-range optics and mirrors near the face. With MEG systems, no auxiliary electromagnetic field may be introduced, and therefore long-distance eye-trackers are also used.

If possible, have your laboratory close to a participant population, or at least make it easy for your participants to reach your lab. That makes it easier to set up a 'production line' where participants arrive one after the other to large recordings.

2.4.2 Types of laboratories and their infrastructure

There are labs that have done eye tracking for 20 years or more, and there are others that have just started. There are labs that only serve a few researchers around the owner of the system, and there are labs that actively invite others to use their equipment, against a cost. There are some eye-tracking companies that conduct studies on a commercial basis.

The largest *commercial* eye-tracking labs have 20–50 eye-trackers to test advertisement campaigns. They are connected to, and sometimes part of, the largest, well-known consumer

product companies, and have gathered the necessary technical and scientific competence in their groups. Unfortunately, they do not publish their work, and they are reluctant to talk about how they use eye tracking and how they are structured. The smallest commercial eye-tracking labs are often media consultants, consisting of one or two people, who often have no previous experience in any of the competence areas necessary to do high-quality eye-tracking work.

Many *academic* eye-tracking laboratories consist of a professor and one or two graduate students and/or post-doctoral researchers, who between them can mostly provide the *scientific* competence needed in their own studies, and who can—if needed—also read up on previous eye-movement research and its technicalities. Some labs quickly grasp the technology of their eye-tracker. If the research group is less technically inclined, the necessary *technical* maintenance is often thought to be a task for the computer technician in the department, the one who is also responsible for email, servers, web, and some programming. However, unless the computer technician learns how to operate and design studies with the eye-tracker, such a division of competences is, in our experience, an unfortunate organization of labs. It typically forces the graduate students to take full charge of the eye-tracker, solve technical issues, upgrade the software, and maintain contact with the manufacturer's support line. The graduate students do this because anyone who makes a change in hardware or the settings of the recording system must also understand that system well enough to be able to make a recording, and see that the data quality requirements for the next study in line are satisfactorily met. Since data quality issues are central throughout the research process, from data recording, over the various stages of data analysis, to the responses from reviewers to submitted manuscripts, satisfactory diagnosis and maintenance of an eye-tracker can only be done by a person confident in *all* aspects of this research process. It can be difficult to find an employee who is sufficiently competent in every one of these skills, and inevitably mistakes will be made as graduate students learn.

Ideally, a *larger lab* is headed by a person who has both technical and research background, someone who can bridge the competence gap that originates from the time when eye-trackers began to be manufactured and sold. This means knowing the recording technology in enough detail to know what a good signal is, to diagnose and remedy errors, to be able to record and analyse data, and follow the research process all the way from hypothesis formulation to reviewer comments and publication.

Since recording high-quality eye-tracking data requires expertise that they can only get from experience, it is important for the staff doing the actual recording work to take part in many recordings with different participants, stimuli, and tasks. As the quality of recorded data is important for subsequent data analysis, it is easier if the same person does both recording and analysis; it is better if the researcher with the highest incentive to get good data takes part in the recordings, so she can influence the many choices made during eye-camera set-up and calibration (pp. 116–134). The exception is when the analysis is subjective in nature and needs to be performed by a person naive to the purpose of the experiment. Any staff who meet and greet participants should have appropriate professionalism for the job; they should be able to answer questions relating to the experimental procedure which the participant is about to undergo.

It is very useful if laboratory staff are also knowledgeable in programming, both for stimulus presentation and for data coding/analysis. Matlab, R, and Python are the preferred software in our and many other labs. If you have scientific ambitions—following the standards of peer-reviewed journals, rather than having heat maps as deliverables—it is also very useful to have a dedicated methodological and statistical specialist in your laboratory.

Since it may be difficult to find all these qualities in one person, you may need several staff members in your lab. Finally, whether you alone carry the full responsibility of your

eye-tracking laboratory, or you share it with others, it is very useful to be part of a laboratory network, sharing experiences, knowledge, and software.

2.5 Measuring the movements of the eye

This section introduces the major eye-movement measuring method in use today, the pupil-and-corneal-reflection method. To better understand the principles of this measurement technique, we will begin with a very brief survey of the eye, and its basic movements.

2.5.1 The eye and its movements

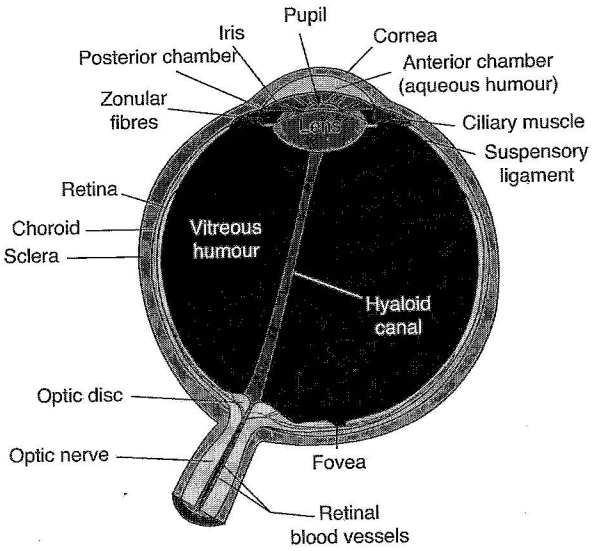
The human eye lets light in through the *pupil*, turns the image upside down in the lense and then projects it onto the back of the eyeball–the *retina*. The retina is filled with light-sensitive cells, called *cones* and *rods*, which transduce the incoming light into electrical signals sent through the optic nerve to the visual cortex for further processing. Cones are sensitive to what is known as high spatial frequency (also known as visual detail) and provide us with colour vision. Rods are very sensitive to light, and therefore support vision under dim light conditions.

There is a small area at the bottom of Figure 2.2(a), called the *fovea*. Here, in this small area, spanning less than 2° of the visual field, cones are extremely over-represented, while they are very sparsely distributed in the periphery of the retina. This has the result that we have full acuity only in this small area, roughly the size of your thumb nail at arm's distance. In order to see a selected object sharply, like a word in a text, we therefore have to move our eyes, so that the light from the word falls directly on the fovea. Only when we *foveate* it can we read the word. Foveal information is prioritized in processing due to the cortical magnification factor, which increases linearly with eccentricity, from about 0.15°/mm cortical matter at the fovea to 1.5°/mm at an eccentricity of 20° (Hubel & Wiesel, 1974). As a result, about 25% of visual cortex processes the central 2.5° of the visual scene (De Valois & De Valois, 1980).

For video-based measurement of eye movements, the pupil is very important. The other important, and less known, element on the eyeball is the *cornea*. The cornea covers the outside of the eye, and reflects light. The reflection that you can see in someone's eyes usually comes from the cornea. When tracking the eyes of participants, we mostly want only one reflection (although in some systems two or more reflections are used), so we record in infrared, to avoid all natural light reflections, and typically illuminate the eye with one (or more) infrared light source. The resulting *corneal reflection* is also known as 'glint' and the '1st Purkinje reflection' (P1). One should also be aware that light is reflected further back as well—both off the cornea and the lens—as illustrated in Figure 2.2(b). The corneal reflection is the brightest, but not the only reflection.

Human eye movements are controlled by three pairs of muscles, depicted in Figure 2.3. They are responsible for horizontal (yaw), vertical (pitch), and torsional (roll) eye movements, respectively, and hence control the three-dimensional orientation of the eye inside the head. According to Donder's law (Tweed & Vilis, 1990), the orientation uniquely decides the direction of gaze, independent of how the eye was previously orientated. Large parts of the brain are engaged in controlling these muscles so they direct the gaze to relevant locations in space.

The most reported event in eye-tracking data does not in fact relate to a movement, but to the state when the eye remains still over a period of time, for example when the eye temporarily stops at a word during reading. This is called a *fixation* and lasts anywhere from some tens



(a) The human eye (From Wikimedia Commons).

(b) The four Purkinje reflections resulting from incoming light.

Fig. 2.2 For eye tracking, the important parts in the order encountered by incoming light are: the *cornea*, the *iris* and *pupil*, the *lens*, and the *fovea*.

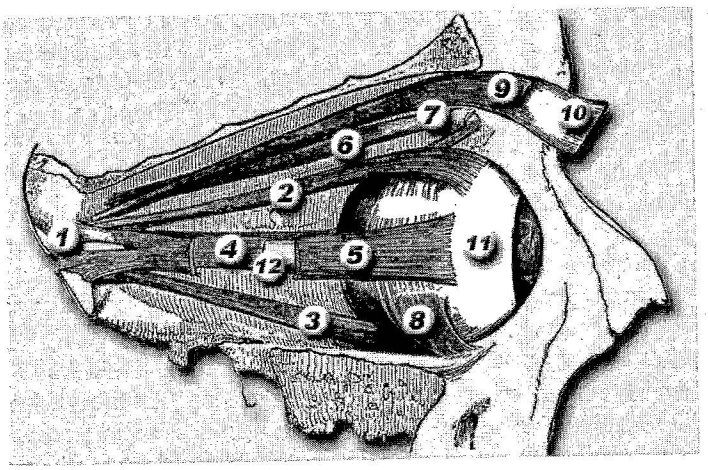


Fig. 2.3 The human eye muscles. The muscle pair (2)–(3) generate the vertical up-down movements, while (4)–(5) generate horizontal right-left movements. The pair (7)–(8) generate the torsional rotating movement. (9)–(10) control the eyelid. Reprinted from Gray's Anatomy of the Human Body, Henry Gray, Copyright (1918).

of milliseconds up to several seconds. It is generally considered that when we measure a fixation, we also measure attention to that position, even though exceptions exist that separate the two.

The word 'fixation' is a bit misleading because the eye is not completely still, but has three distinct types of micro-movements: *tremor* (sometimes called physiological nystagmus), *microsaccades* and *drifts* (Martinez-Conde, Macknik, & Hubel, 2004). Tremor is a small movement of frequency around 90 Hz, whose exact role is unclear; it can be imprecise

Table 2.3 Typical values of the most common types of eye movement events. Most eye-trackers can only record some of these.

Type	Duration (ms)	Amplitude	Velocity
Fixation	200–300	–	–
Saccade	30–80	4–20°	30–500°/s
Glissade	10–40	0.5–2°	20–140°/s
Smooth pursuit	–	–	10–30°/s
Microsaccade	10–30	10–40′	15–50°/s
Tremor	–	$< 1'$	20′/s (peak)
Drift	200–1000	1–60′	6–25′/s

muscle control. Drifts are slow movements taking the eye away from the centre of fixation, and the role of microsaccades is to quickly bring the eye back to its original position. These intra-fixational eye movements are mostly studied to understand human neurology.

The rapid motion of the eye from one fixation to another (from word to word in reading, for instance) is called a *saccade*. Saccades are very fast—the fastest movement the body can produce—typically taking 30–80 ms to complete, and it is considered safe to say that we are blind during most of the saccade. Saccades are also very often measured and reported upon. They rarely take the shortest path between two points, but can undergo one of several shapes and curvatures. A large portion of saccades do not stop directly at the intended target, but the eye 'wobbles' a little before coming to a stop. This post-saccadic movement is called a *glissade* in this book (p. 182).

If our eyes follow a bird across the sky, we make a slower movement called *smooth pursuit*. Saccades and smooth pursuit are completely different movements, driven by different parts of the brain. Smooth pursuit requires something to follow, while saccades can be made on a white wall or even in the dark, with no stimuli at all.

Typical values for the most common types of eye movements are given in Table 2.3. While these eye movements are the ones most researchers report on, especially in psychology, cognitive science, human factors, and neurology, there are several other ways for the eye to move, which we will meet later in the book.

Rather than mm on a computer screen, eye movements are often measured in *visual degrees* (°) or *minutes* (′), where $1° = 60'$. Given the viewing distance d and the visual angle θ, one can easily calculate how many units x the visual angle spans in stimulus space. The geometric relationships between these parameters are shown in Figure 2.4, and can be expressed as

$$\tan\frac{\theta}{2} = \frac{x}{d} \tag{2.1}$$

Note, however, that this relationship holds only when the gaze angle is small, i.e. when the stimulus is viewed in the central line of sight. For large gaze angles, the same visual angle θ_1 may result in different displacements (x_1 and x_2) on the stimulus, as illustrated in Figure 2.4.

If your stimulus is shown on a computer screen, you may want to use pixels units instead of e.g. mm. If $M \times N$ mm denotes the physical size of a screen with resolution $r_x \times r_y$ pixels, then 1 mm on the screen corresponds to r_x/M pixels horizontally and r_y/N pixels vertically.

When measuring eye-in-head movement, visual angle is the only real option to quantify eye movements, since the movements are not related to any points in stimulus space. Visual angle is also suitable for head-mounted eye tracking in an unconstrained environment, e.g. a supermarket, since the distance to the stimulus will change throughout the recording.

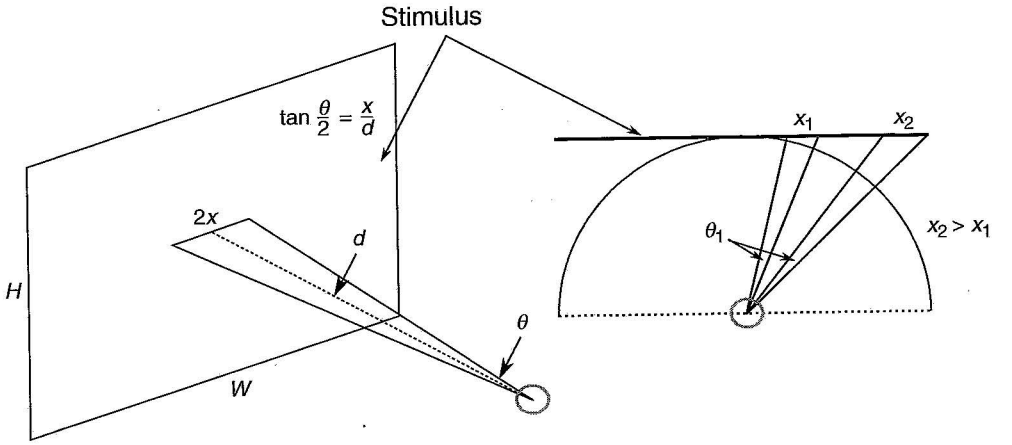


Fig. 2.4 Geometric relationship between stimulus unit x (e.g., pixels or mm) and degrees of visual angle θ, given the viewing distance d. Notice that on a flat stimulus, the same visual angle (θ_1) gives two different displacements (x_1,x_2).

2.5.2 Binocular properties of eye movements

An important aspect of human vision is that both eyes are used to explore the visual world. When using the types of movements defined in the previous section, the eyes sometimes move in relation to each other. *Vergence* eye movement refers to when the eyes move in directly opposite directions, i.e. converging or diverging. These opposite movements are important to avoid double vision (diplopia) when the foveated object moves in a three-dimensional space.

For most participants, both eyes look at the same position in the world. But many people have a dominant eye, and one which is more passive. If the difference is large, the passive eye may be directed in a different direction from that of the dominant one, and we say colloquially that the participant is *squinting*. The technical term is either *binocular disparity* or *disjugacy*.

In reading, disparity can be in the order of one letter space at the onset of a new fixation, and can occur in more than half of the fixations. Liversedge, White, Findlay, and Rayner (2006) found that disparity decreased somewhat over the period of the fixation, but not completely, and was of both crossed (right eye to the left of the average gaze position and vice versa) and uncrossed nature. In a non-reading 'natural' task, Cornell, MacDougall, Predebon, and Curthoys (2003) reported disparities of up to ±2°, but also noticed that disparities of 5° were present (but rare) in the data. All these results were found for normal, healthy participants.

While it is commonly believed that the eyes move in temporal synchrony, binocular coordination varies over time. During the initial stage of a saccade, for example, the abducting eye (moving away from the nose) has been found to move faster and longer than the adducting eye (moving towards the nose) (Collewijn, Erkelens, & Steinman, 1988). At the end of the saccade this misalignment is corrected, both through immediate glissadic adjustments, and slower post-saccadic drift during the subsequent fixation (Kapoula, Robinson, & Hain, 1986).

2.5.3 Pupil and corneal reflection eye tracking

The dominating method for estimating the point of gaze—where someone looks on the stimulus—from an image of the eye is based on pupil and corneal reflection tracking (see D. W. Hansen & Ji, 2009, Hammoud, 2008, and Duchowski, 2007 for technical details and

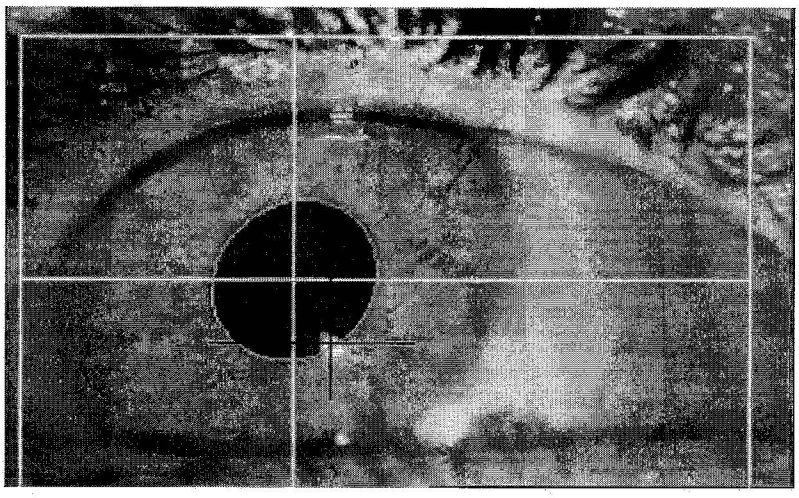


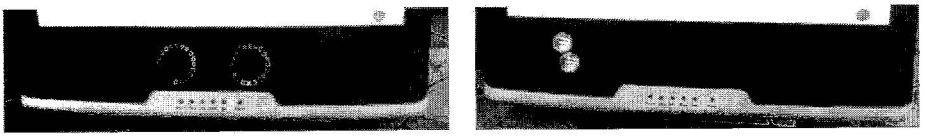
Fig. 2.5 A pupil–corneal reflection system has properly identified the pupil (white cross-hair) and corneal reflection (black cross-hair) in the video image of a participant's eye.

an overview of other methods).

A picture of an eye with both pupil and corneal reflection correctly identified can be seen in Figure 2.5. While it is possible to use pupil-only tracking, the corneal reflection offers an additional reference point in the eye image needed to compensate for smaller head movements. This advantage has made video-based pupil and corneal reflection tracking the dominating method since the early 1990s.

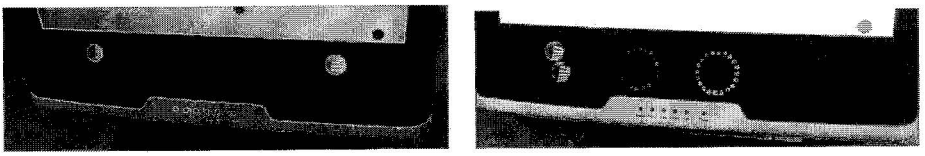
The pupil can either appear dark in the eye image, which is the most common case, or bright, as with some ASL (Applied Science Laboratory), LC Technology, and Tobii systems. The bright pupil is bright because of infrared light reflected back from the retina, through the pupil. Such a system requires the infrared illumination to be co-axial with the view from the eye camera, which puts specific requirements on the position of cameras and illumination (Figure 2.6(a)). As long as the pupil is large, a bright-pupil system operates in approximately the same way as a dark-pupil system, but for small pupil sizes (when there is a lot of ambient light), a bright-pupil system may falter. The original motivation behind bright-pupil systems appears to have been to compensate for poor contrast sensitivity in the eye camera by increasing the difference in light emission between pupil and iris, but with new improved camera technology, contrast between iris and pupil is often also very good for most dark-pupil systems, as can be seen in Figure 2.5. At least one eye-tracker has been built that switches between bright and dark tracking mode, which requires the turning on and off of several infrared illuminators depending on how well tracking works in the current state (Figure 2.6). No studies have systematically investigated which of the two tracking modes gives better data quality over large populations, but in the authors' experience, data quality rather depends on the quality of the eye camera and other parts of the eye-tracker.

Noteworthy is that the methods used for image analysis and gaze estimation can vary significantly across different eye-trackers, both freely available and commercial. Therefore it may be difficult to compare systems between different manufacturers. To complicate the issue even further, some eye-tracking manufacturers keep many key technical details about the system secret from the user community. Sometimes the user is not allowed to see how the eye image is analysed, for instance, but only a very simplified representation of the position of the eyes is given. Figure 2.7 shows the eye image and the simultaneous simplified representation of the eyes, on a *remote system* (p. 51). If the recording software allows the operator to see



(a) Bright pupil mode: A ring of infrared diodes around the two eye cameras, making illumination almost co-axial with camera view.

(b) Single side dark-pupil mode: Diodes off-axis to the left.



(c) Dual side dark-pupil mode: Diodes off-axis on both sides.

(d) Searching: Rapidly switching between dark- and bright-pupil illumination.

Fig. 2.6 Four illumination states of the Tobii T120 dual mode remote eye-tracker. This particular eye–tracker changes to another tracking mode when tracking fails in the current mode.

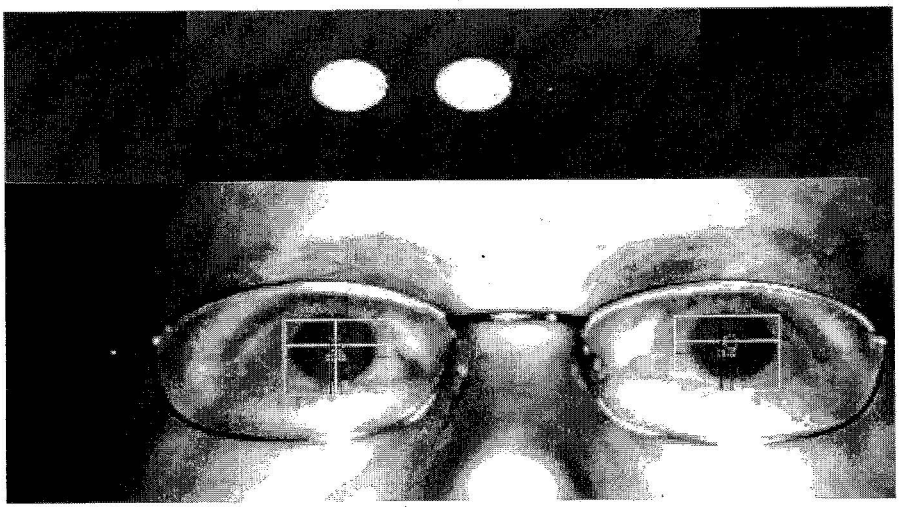


Fig. 2.7 Eye image in bottom half; and simplified representation of the eyes at the top.

the eye image, it is easier to set up the eye camera to ensure that tracking is optimal. Access to the eye image also makes it easier to anticipate and detect potential problems before and during data collection (pp. 116–134).

Figure 2.8 shows a schematic overview of a video-based eye-tracker, where the operations required to calculate where someone looks have been divided in three main blocks: *image acquisition*, *image analysis*, and *gaze estimation*.

In the acquisition step, an image of the eye is grabbed from the camera and sent for analysis. This can usually be done very quickly, but if head movement is allowed (as in remote eye-trackers), the first step of the analysis is to detect where the face and eyes are positioned in the image, whereafter image-processing algorithms segment the pupil and the corneal reflection from a zoomed-in portion of the eye. Geometrical calculations combined with a calibration procedure are finally used to map the positions of the pupil and corneal reflection to the data sample (x, y) on the stimulus.

While the pupil is a part of the eye, the corneal reflection is caused by an infrared light

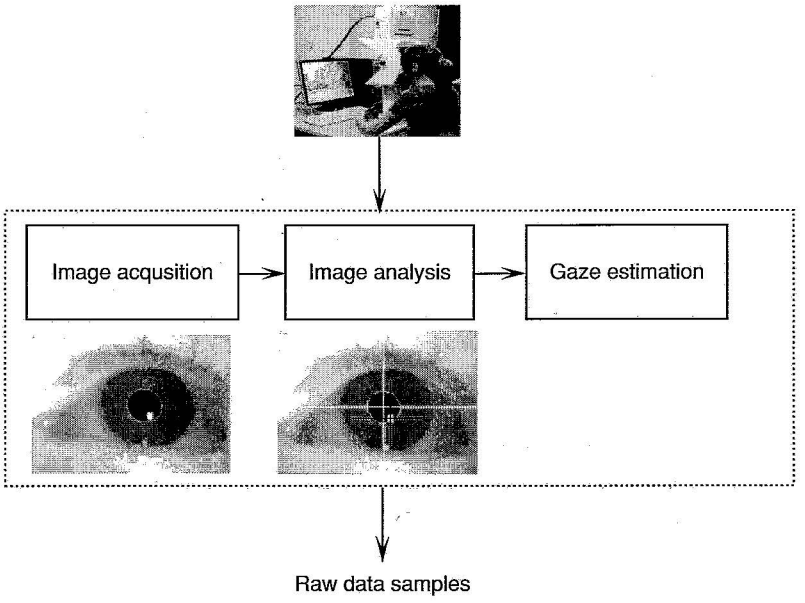


Fig. 2.8 Overview of a video based eye-tracking system.

source positioned in front of the viewer. The overall goal of image analysis is to robustly detect the pupil and the corneal reflection in order to calculate their geometric centres, which are used in further calculations. This is typically done using either *feature-based* or *model-based* approaches. The feature-based approach is the simplest where features in the eye image are detected by criteria decided automatically by an algorithm or subjectively by the experimenter. One such criterion is thresholding, which finds regions with similar pixel intensities in the eye image. Having access to a good eye image where the pupil (a dark oval) and the corneal reflection (smaller bright dot) are clearly distinguishable from the rest of the eye is important for thresholding approaches. Another feature-based approach looks for gradients (edges, contours) that outline regions in the eye image that resemble the target features, e.g. the pupil.

To increase the precision of the calculation of geometric centres, the algorithms typically include sub-pixel estimation of the contours outlining the detected features. The principal calculation is illustrated in Figure 2.9.

The major weakness of feature-based pupil–corneal reflection systems is that the calculation of pupil centre may be disturbed by a descending eyelid and downward pointing eye lashes. Lid occlusion of the pupil may cause—as we will see on pages 116–134 on camera set-up—*offsets* (incorrectly measured gaze positions) and increased *imprecision* in the data in some parts of the visual field. Figure 2.10 shows a participant with a drooping eyelid and downward eyelashes. The pupil is covered and cannot be identified, while the corneal reflection is dimly seen among the lashes.

A second weakness of feature-based systems concerns extreme gaze angles, at which the corneal reflection is often lost, but as we explain on pages 116–134, this can often be solved by moving the stimulus monitor, eye cameras, or infrared sources.

A third and mostly minor weakness is that the measured gaze position may be sensitive to variations in pupil dilation. In recordings where accuracy errors are not tolerated—as in the control systems for the lasers used in eye surgery—another technology called limbus-tracking is used. The limbus is the border between the iris and the sclera. It is insensitive to variations

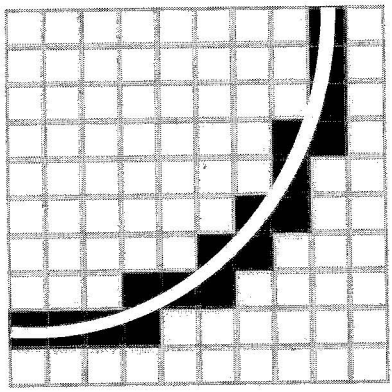


Fig. 2.9 Increasing precision by sub-pixel estimation of contours.

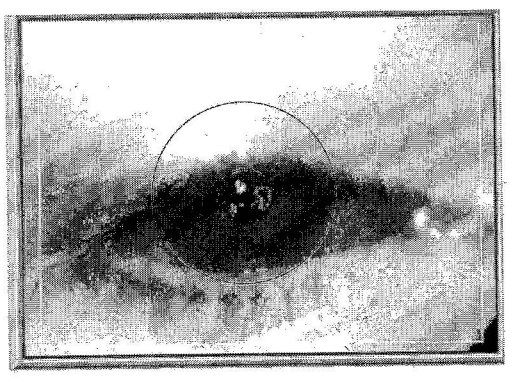


Fig. 2.10 In model-based eye tracking, the recording computer uses a model of the eye to calculate the correct position of the iris and the pupil in the eye video, even if parts of them are occluded by the eyelids. Rings indicate features of the model.

in pupil dilation, but very sensitive to eyelid closure, and is therefore fairly impractical except in specific applications such as laser eye surgery.

Model-based solutions can alleviate the weakness of feature-based pupil–corneal reflection systems by using a model of the eye that fits onto the eye image using pattern matching techniques (Hammoud, 2008). A useful eye model would assume that both iris and pupil are roughly ellipsoidal and that the pupil is in the middle of the iris. For example, using an ellipsoidal fit of the pupil would prevent the calculated centre of the pupil moving downward when the upper eyelid occludes the top part of the pupil, something that happens in the beginning and at the end of each blink, or when the participant gets drowsy.

Another model assumption is that the pupil is darker than the iris, which is darker than the sclera. When the model knows this, it can position the iris and pupil circles onto the most probable position in the image.

Moriyama, Xiao, Cohn, and Kanade (2006) implemented and tested a model-based iris detector that could be used for eye tracking. Although it does solve many eyelid problems known from feature-based pupil–corneal reflection systems, the model-based eye-tracker still sometimes misplaces the iris due to shadows in the eye image. While having the potential to provide more accurate and robust estimations of where the pupil and corneal reflection centres are located, model-based approaches add significantly to the computational complexity since they need to search for parts of the eye image that best fit the model. Without a good initial

guess of where the pupil and corneal reflections are located, the time it takes the algorithm to find the indented features (often called *recovery time*) may be unacceptably long. Fortunately, a full search is needed only in the first frame, since feature positions in subsequent frames can be predicted from previous ones. Note that recovery time is very individual, since the algorithm will find some participants' eyes much faster than others. As feature-based approaches can provide this first guess, eye-tracking approaches that combine feature- and model-based approaches will probably become even more common in the future.

Now assume that the centres of both pupil and corneal reflection have been correctly identified, that the head is fixed and that we have a complete geometrical model of the eye, the camera, and the viewing set-up; then the gaze position could be calculated mathematically (Guestrin & Eizenman, 2006). However, this is usually not done in real systems, mainly due to the difficulty in obtaining robust geometric models of the eye. Instead, the majority of current systems use the fact that the relative positions of pupil and corneal reflections change systematically when the eye moves: the pupil moves faster, and the corneal reflection more slowly. The eye-tracker reads the relative distance between the two and calculates the gaze position on the basis of this relation. For this to work, we must give the eye-tracker some examples of how points in our tracked area correspond to specific pupil and corneal reflection relations. We tell this to the eye-tracker by performing a *calibration*, which typically consists of 5, 9, or 13 points presented in the stimulus space that are fixated and sampled one at the time. The practical details of calibration are described on pages 128–134.

While using one camera and one infrared source works quite well as long as the head is fairly still, more cameras and infrared sources can be used to relax the constraints on head movement and calibration. Using two infrared sources gives another reference point in the eye image, and is in theory the simplest system that allows for free head movement, which is desirable in remote eye-trackers. Using multiple cameras and infrared sources, it is theoretically possible to use only one point to calibrate the system (Guestrin & Eizenman, 2006). However, using another light source complicates the mathematical calculations.

The most common commercially available eye-trackers are those with one or two cameras and one or multiple infrared light sources that work best with 5, 9, and 13 point calibrations. More on different types of eye-camera set-ups can be found on pages 51–64.

2.6 Data quality

Data quality is a property of the sequence of raw data samples produced by the eye-tracker. It results from the combined effects of eye-tracker specific properties such as sampling frequency, latency, and precision, and participant-specific properties such as glasses, mascara, and any inconsistencies during calibration and in the filters during and after recording.

Some eye-trackers also output pupil and corneal reflection positions, calculated directly from the eye image prior to gaze estimation. We will not talk about the quality of this position data as they are not used very often; however, the reader should observe that they have properties common to the sequence of raw data samples.

Data quality is of the utmost importance, as it may undermine or completely reverse results. Already McConkie (1981) argues that every published research article should list measured values for the quality of the data used, but this has not yet become the standard.

2.6.1 Sampling frequency: what speed do you need?

The sampling frequency is one of the most highlighted properties of eye-trackers by manufacturers, and there is a certain competition in having the fastest system. You do need some

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 J. M. Henderson and Hollingworth (1999), J. M. 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* (J. 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.

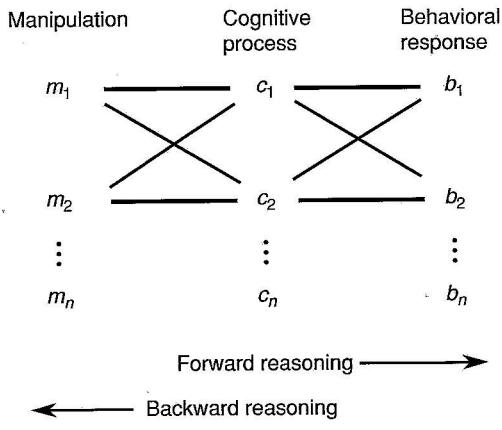


Fig. 3.1 Available reasoning paths: possible paths of influence that different variables can have. Our goal is to correctly establish what variables influence what. Notice that there is a near-infinite number of variables that influence, to a greater or lesser degree, any other given variable.

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 Y. 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.

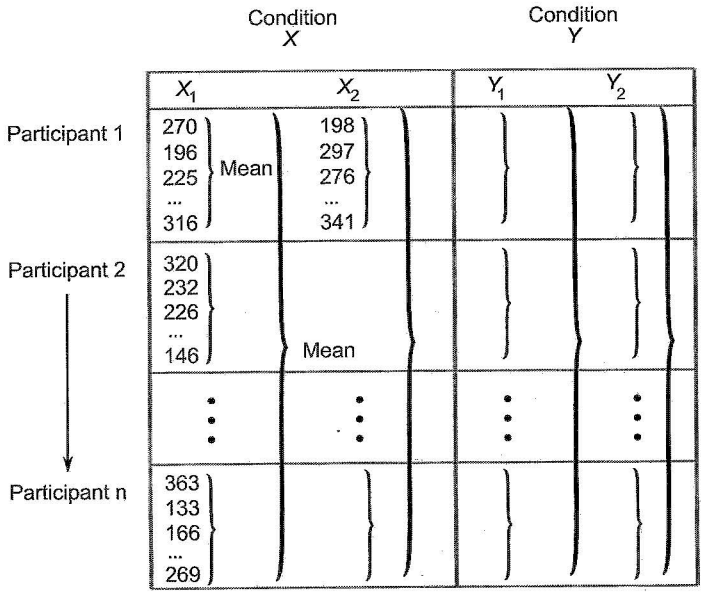


Fig. 3.2 A typical experimental design, and how means are calculated from it. Here we have two independent variables, X and Y, each with two factors, $X_{1,2}$ and $Y_{1,2}$. The conditions could be preferred and non-preferred music, each with high and low volume level, for instance. Each number represents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

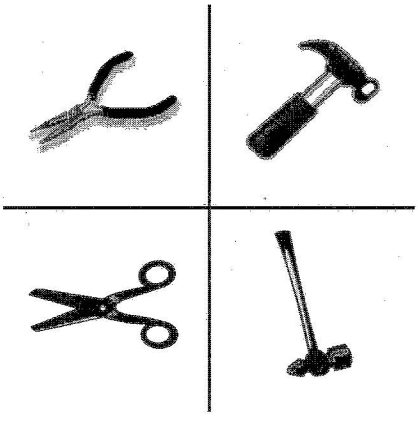
When selecting the precise stimuli, it is useful to consider what it is that generally draws our attention, so the effect of primary interest is not blocked or completely dominated by other larger effects. Below are a few examples of factors that are known to influence the allocation of visual attention,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

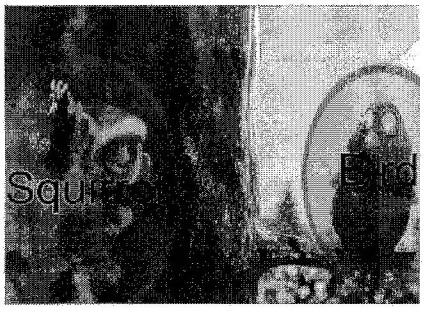
Stimulus images are often divided into AOIs, the 'areas of interest', which are sometimes also called 'regions of interest'. How to make this division is discussed in depth in Chapter 6. In short, the researcher chooses AOIs while inspecting potential stimulus pictures with the precise hypothesis and 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

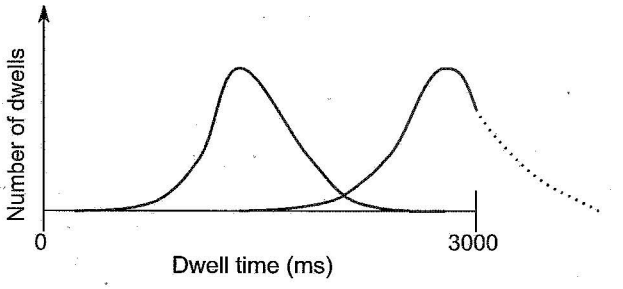


Fig. 3.4 Dotted line of one distribution of dwell times is outside of the fixed trial duration.

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

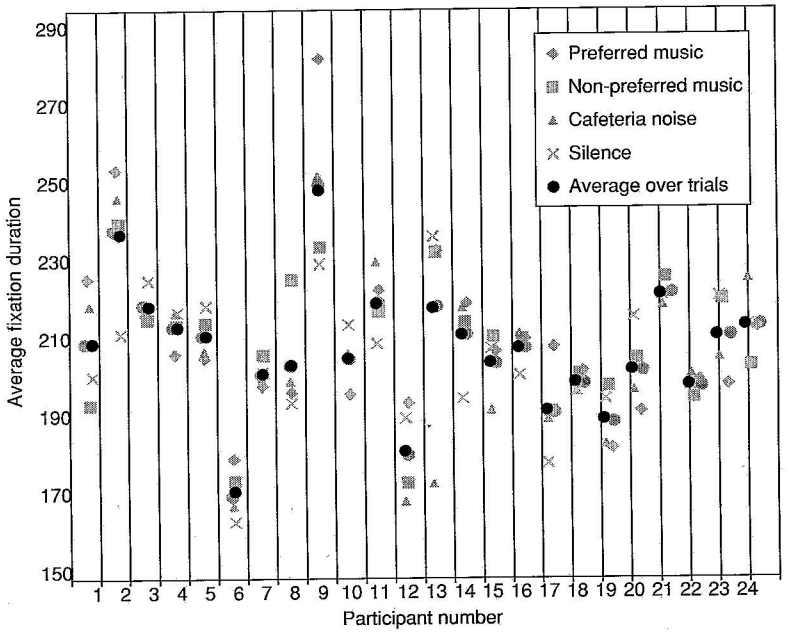


Fig. 3.5 Idiosyncracy: every participant has his or her own individual average fixation duration and exhibits it across different recordings. Individual participant variation is large, and the effect of the experimental conditions is small.

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

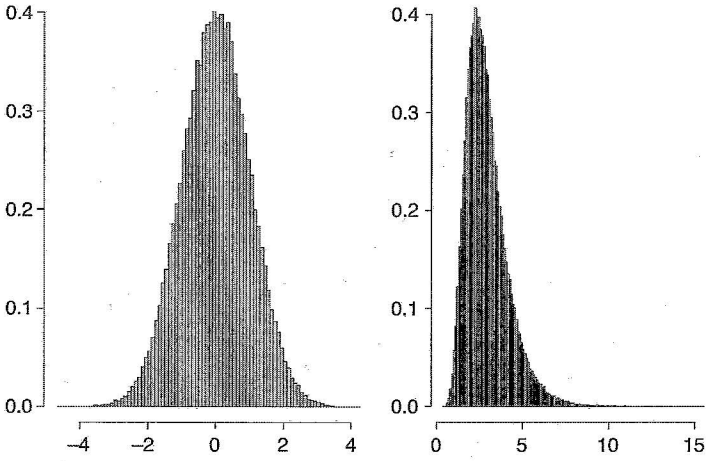
was carried out. Feeding the data into a data analysis without checking for errors may have devastating effects, either producing significant effects that do not exist, or hiding them.

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.

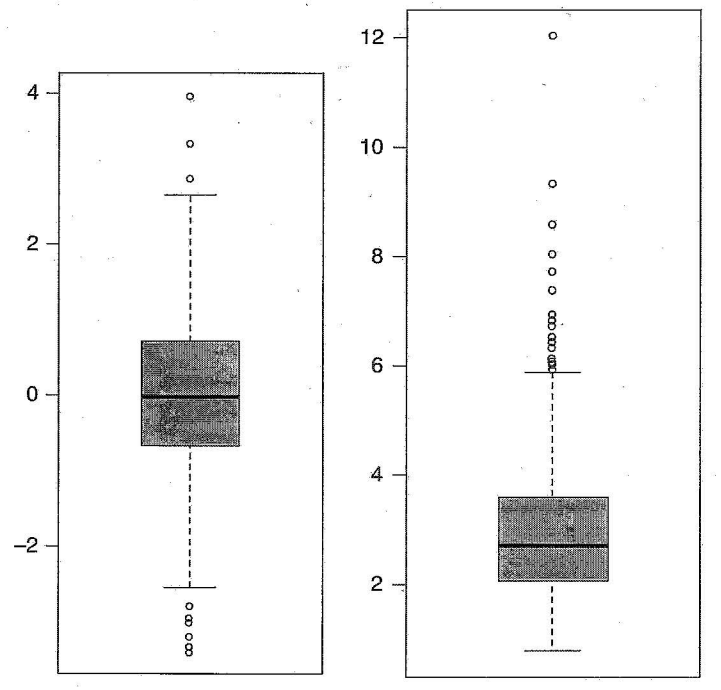
Plotting is also an indispensable tool in the later stages of data exploration. Particularly useful are box-and-whiskers plots, which give simultaneous information about the distribution as well as potential outliers (compare Figure 3.7). Additional plots that might be helpful are histograms (as in Figure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

experimenter to apply to a regional board of ethics before running the study. The ethical board will then decide whether the study follows the local law, and whether the scientific purpose of the study outweighs the hardships and/or sufferings of the participants. You may think that participants do not suffer very much from taking part in an eye-tracking study, but participants with clinical diagnoses, such as dyslexia, can feel bad about doing a reading study in a highly technical laboratory, in particular if the participant is young and has already been subjected to excessive testing.

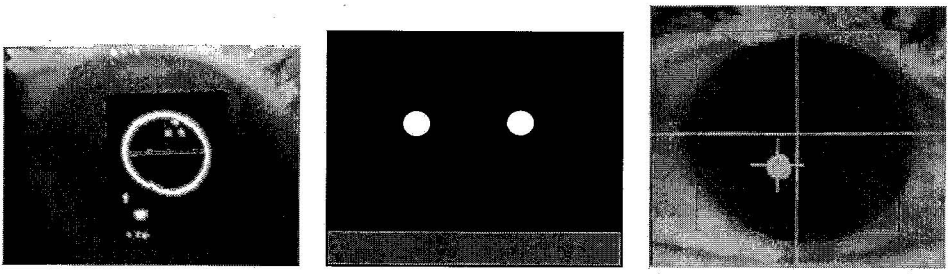
Ethics fundamentally means that the participant should have a good feeling about having participated in your study. This is in your own interest, because your participant will tell his friends about your experiment and the laboratory in which it took place. Not only will you increase the chances that your participant will come back therefore, you potentially gain new participants for future studies also. If you have treated your participants well, and if they are interested in your study, they may even recruit their friends for you! If you get them to be really interested, after a few years, one of your former participants may suddenly be your student. All in all, you should treat your participants well.

Many ethics-related issues are dealt with in the 'consent form', a written statement that the researcher may use the data collected from the participant, and that the participant considers himself fully informed about the purpose of the study and the consequences for himself of signing off the rights to the data. Laws may require consent forms to be signed *before* data recording, but the objectives of the study sometimes require this to be done *after* the recording (so that the participant is not aware of the purpose of the study during the recording). This can be solved by preparing two consent forms: one which is more general in its experiment description, and a second post-experiment consent form which fully explains the experiment and gives the participant the option to have the recorded data destroyed if no consent is given.

4.4 Eye camera set-up

Now we are ready to put the participant into one of the eye-trackers described in Chapter 2. This section describes how to adjust the eye camera view so as to get data of high quality. Eye camera set-up is of great importance to the quality of your data, and largely decides whether you can use them in analysis, and for what measures. Knowing your eye video and the algorithms that calculate gaze position from it allows you to get better data from a larger spectrum of participants, more quickly and with less anxiety.

Some manufacturers have hidden the eye camera set-up in automatic processes, giving an appearance of increased usability and simplicity (Figure 4.1(b)), but at the same time withholding control over the data recording from the user of the system. The result in terms of data quality may be acceptable in a large number of cases, but some loss of quality will go unnoticed, or be difficult to understand and alleviate. Experienced users of these mostly remote systems will with time learn to move or tilt a whole system (camera unit and monitor) to set up an optimal eye-camera angle, behaving in accordance with the advice we give below, but without the feedback that a full eye-image can provide. Most eye-tracking recording software shows one or another image of the participant's eye, however. Figures 4.1(a) and (c) are examples from the ASL MobileEye, and the SR EyeLink respectively. In this section, we will use examples from SMI and EyeLink systems, because the video image can be very clearly seen with their systems, but the presentation and majority of advice in this section are valid for all video-based pupil–corneal reflection systems. We will use reading data to show the effect on data quality, because the lines of text provide a good reference point against which to compare data.



(a) The ASL MobileEye eye image (b) The Tobii Studio eye image (c) The EyeLink 1000 eye image

Fig. 4.1 The eye images of three video-based pupil-corneal-reflection eye-tracking systems. To the left ASL MobileEye eye image, in which the highlighted ring indicates the circumference of the pupil. In the middle, Tobii Studio in which the eye image is hidden, and to the right the EyeLink eye image, with pupil and corneal reflections clearly marked by colour and cross-hairs.

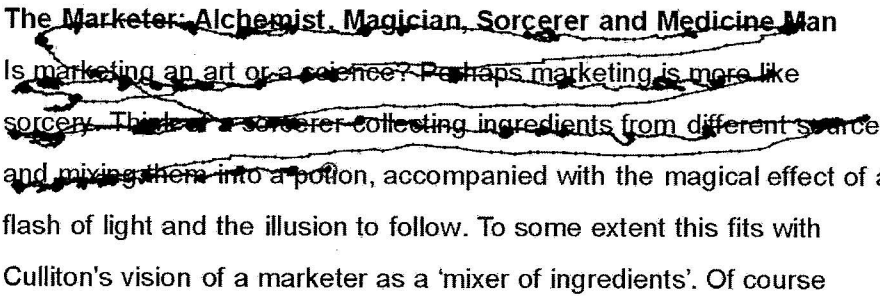


Fig. 4.2 Example of good quality data in raw format, recorded with a tower-mounted eye-tracker at 1250 Hz. The figure shows raw gaze samples; one dot per sample. Fixations are seen as blobs of many dots in a small area. Saccades are strings of dots; the more sparsely the dots are placed, the faster the saccade.

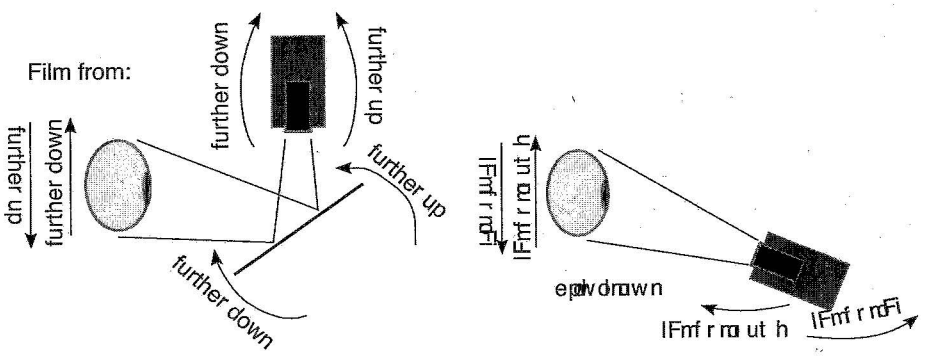
Errors can arise in the eye (feature) detection and gaze estimation algorithms at the heart of the eye-tracker due to *covering* of the pupil or corneal reflection, *confusion* between the pupil or corneal reflection and a number of other optic entities, *distortion* of the image, or because of *loss* either of the pupil or corneal reflection by the camera and its supporting software. A number of causes behind these problematic conditions are summarized in Table 4.1.

In order to get high-quality data such as that shown in Figure 4.2, you should see to it that the eye camera records the eye from slightly below the eye with regard to the direction the participant has when facing the stimulus area. If your eye camera is too low, it will be difficult to calibrate and record in the upper corners of the calibration area because the corneal reflection is more easily occluded by the lower eyelid. If the camera angle is too high, the upper eyelid will often cover some participants' pupils when looking at the bottom corners, and calibration will be difficult or faulty. Figure 4.3 shows how camera positions, the participant, and the mirror should be moved to achieve a better image. When you are moving cameras or mirrors close to your participant's eyes, take care always to talk to him about what you are doing, so that he is prepared and you do not startle him. Remember that by participating, he puts trust in you. Of course, you must be careful not to hurt him or scratch his glasses.

The participant and cameras should be placed in a position so that the eye camera shows a good image of the eye for each corner of the calibration area. This involves directing cameras and/or mirrors as well as raising or lowering the participant's chair, or the table with eye-

Table 4.1 Summary of optic conditions that may endanger data quality. For instance, both pupil and corneal reflection may be covered; the pupil foremost by droopy eyelids, and the reflection by laughter, or more specifically, the accompanied narrowing or closure of the eye. Confusion refers to cases when other objects are mistaken for the whole or part of the pupil or corneal reflection. Optic distortion can be caused by, for instance, bifocal glasses, while extreme gaze angles may cause loss of the corneal reflection.

	Pupil	**Reflection**
Covering	Droopy eyelids	Laughter
Confusion	Mascara	Retinal
	Glasses	Glasses
	Ambient infrared light	Contact lenses
	Retinal reflection	Sunlight
	Specks and dirt	Lamps
		Other infrared light sources
		Wet eye
Distortion	Bifocal glasses	
Loss	Head movement	Extreme gaze angles



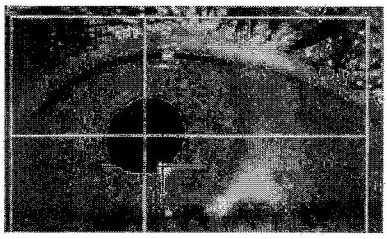
(a) Tower-mounted and some head-mounted systems

(b) Remote systems and some head-mounted systems

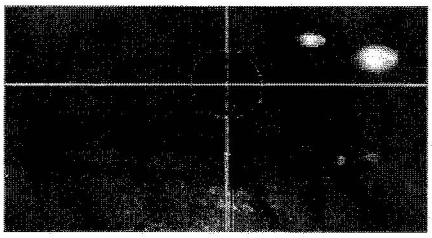
Fig. 4.3 How to move the eye camera, the mirror, or the participant, so as to get a camera angle from further up or further below.

tracker and stimulus monitor. If you are using a head-mounted system, camera set-up often involves positioning and directing a whole construction of mirror and cameras. If you are using a remote eye-tracker or a head-restraining static system, remember that participants tend to sit very attentively during camera set-up and calibration, but may slide down during the actual recording, so your camera angle changes or you have to move mirrors or camera during recording (or it is done automatically), which may cause an offset in data.

A pupil–corneal reflection system calculates the participant's gaze direction on the basis of the relative position of the pupil and the corneal reflection centres. When calculating the centre of the pupil, it is important that the eye camera sees the entire pupil in all gaze directions. Two examples of good eye images can be seen in Figure 4.4. Partial occlusion of the pupil causes a movement of the calculated centre point, and thus a movement in the data coordinates, although no real eye movement has been made. In fact, even very small movements of the calculated pupil centre can cause considerable movement in the coordinate files. The identified corneal reflection may also move artificially, jumping to other reflections



(a) Both pupil and corneal reflection are uniquely identified with white and dark cross-hairs, respectively.



(b) The large white blobs in the right-hand picture are infrared reflections in the participant's glasses, but they do not interfere with the gaze estimation algorithm.

Fig. 4.4 Eyes with good eye camera set-up, filming from below. Both eye images are from head-mounted systems.

in the eye video, most commonly just under or on top of the upper eyelid. This movement of the calculated corneal reflection causes large and very fast, but entirely false, movements in the data samples, which we call *optic artefacts* in Chapter 2.

Ocular dominance

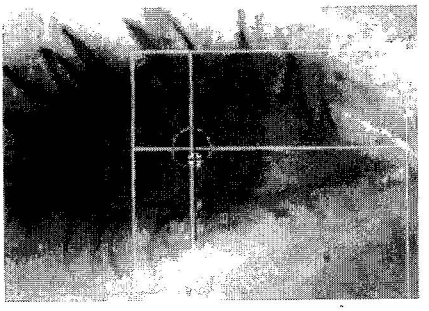
If your participant is squinting or if there is reason to believe that one eye is better than the other, see to it that you select the so-called *dominant* eye. Ocular dominance can be determined using a number of different tests; one simple test for this is the 'Miles test' (Miles, 1929). Here, the participant extends both arms and brings both hands together to create a small opening while looking at an object with both eyes simultaneously. The observer then alternately closes the left and the right eye to determine which eye is really viewing the object. That is the eye that should be measured. Around two thirds of the population have a right eye dominance and men are more commonly right dominant (Eser, Durrie, Schwendeman, & Stahl, 2008).

If you are recording monocularly and you notice that you have to change the camera to the other eye, for instance recording a left eye after a series of right eyes, you may need to take care that the position of the left eye relative to the monitor will be the same as that of the previously recorded right eyes (if your experiment demands control over this). If not, the result may be small variations in fixation and saccade data in the outskirts of the monitor. The best solution is to control for this in the experimental design, so horizontal stimuli positions are counterbalanced.

4.4.1 Mascara

Mascara is considered a serious problem for data quality. This may be less of an issue for a bright-pupil system, as the pupil is bright but the mascara dark (see Applied Science Laboratories, 2011), but mascara also blocks filming through sparse eyelashes, making it not completely unproblematic for bright-pupil systems. Therefore, regardless of eye-tracking system, always tell your participants beforehand not to wear mascara. Do not tell them not to wear make-up, because some participants think that mascara does not qualify as make-up. If the mascara is left on, only expect to make a good recording on participants with large, open eyes and upward eyelashes. This is because in a dark-pupil system, the software that identifies the pupil is confused by the other large dark area in the immediate vicinity of the pupil, and may lock onto the mascara rather than the pupil. This is detrimental to subsequent calibration, as well as to the recording. A make-up removal kit should therefore be an indispensible piece of

(a) Participant with drooping eyelid and downward eyelashes. A thick, dark brush of lashes melds with the pupil and makes it impossible for the recording software to identify the pupil.



(b) Another participant with mascara, but with upward-pointing lashes; in this case the mascara is easy to exclude, and lashes like these seldom occlude the pupil.

Fig. 4.5 Mascara interferes with pupil detection.

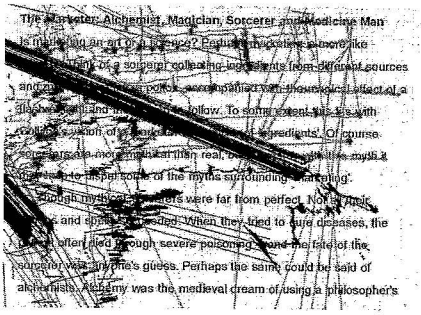
equipment in all serious eye-tracking laboratories. If you ask your participant to remove the mascara, watch out for any remaining mascara blobs that often have a round pupil-like form which can be mistaken for a pupil and later cause offsets in your data.

Some recording software can exclude portions of the eye video from the eye feature detection process, which allows quality recordings without asking participants to remove mascara. In Figure 4.5, only the parts of the eye video that are inside the white rectangle are analysed for pupil and corneal reflection. This works relatively well in the majority of cases, provided the lashes point upwards and the head is fixed so the eye does not move relative to the eye image (as for a fixed contact system with chin rest). Moreover, software programs often give the user the option to reject pupil-like objects that are too small or too large, such that mascara blobs with extreme sizes can be discarded automatically.

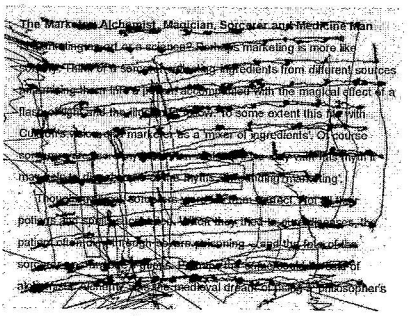
The right side of Figure 4.6 shows a raw, unfiltered plot of data samples from a participant for whom we had a good eye video in all gaze directions. It shows one pixel for each sample, of which there are 1250 per second. Where samples cluster to blobs, the eye has been still in a fixation. Where samples trace a line, the eye has moved quickly in a saccade. The vertical pairs of lines are blinks; the eyelid going down and then up again. All else is noise of various kinds. It can been seen that we have fixations and saccades almost right on the word for each line. The left side of Figure 4.6 shows data from a participant with mascara and downward eyelashes. The data plot is largely OK all the way to line eight, which is the gaze direction at which the mascara aligns with the pupil in the eye video. Data for the lower five lines consist of useless optic artefacts.

4.4.2 Droopy eyelids and downward eyelashes

The problem with droopy eyelids grows with the age of your participants, but it is also a matter of individual variation. Droopy eyelids are a major problem in eye-tracking research even if there is no mascara, because the eyelids, or the eyelashes, cover the pupil in the lower gaze directions (Figures 4.7 and 4.8). In your data, you will see very large downward offsets, because a pupil partially occluded from above has a lower mass centre and resulting data will be at an artificially lower vertical position. At a certain gaze angle, the pupil is completely covered, and you will then have complete data loss in the lower part of the screen/visual field. Notice that pupil occludance may not appear dangerous during calibration, when the participant is more tense and focused, and the eyes are more open. It is when he gets into



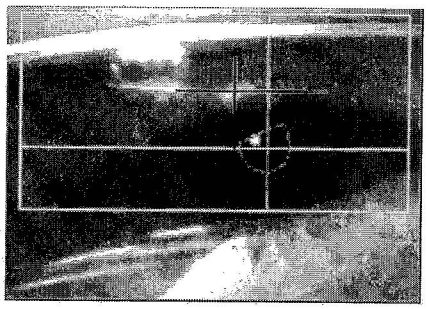
(a) Data from participant wearing mascara and having slightly downward eyelashes. Data plotted in the upper half of the stimulus image shows data recorded when the eye was open and all of the pupil seen. Below this is the data where the pupil is covered by lashes with mascara.



(b) Good eye-tracking data. Vertical lines are blinks. Some noise can be seen in the upper left corner and the bottom line. However, all lines of text have accurate and precise data samples on top of them, that the event detection algorithms can make use of.

Fig. 4.6 The effects of mascara on slightly downward eyelashes.

(a) A thick brush of lashes covers the pupil and make it impossible for the recording software to identify either pupil, or corneal reflection.



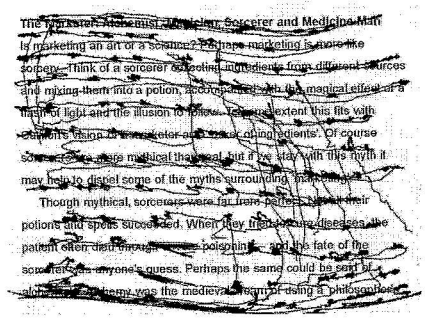
(b) In this drooping eyelid, on a participant with glasses, the eye-tracker locks onto a false corneal reflection on the eyelid, because the real corneal reflection is partially occluded underneath lashes. Also notice the partial pupil occlusion. Gaze data from this eye video will have considerable data loss, a very large offset, and many optic artefacts.

Fig. 4.7 Participant with drooping eyelid and downward eyelashes.

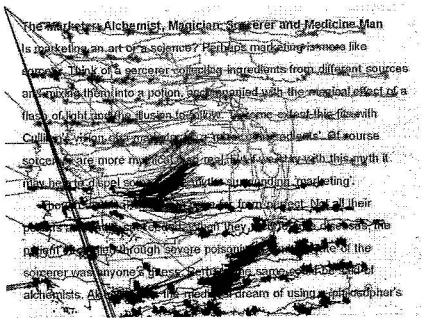
the task and relaxes that the eye closes and you get an offset. In fact, you may get better data quality if you calibrate the bottom calibration points with a relaxed participant, who closes his eyes a little, because that is what the eye will look like when you record your participant looking in that direction. However, that is a gamble and the correct solution is to redo the set-up to handle the drooping eyelids or the eyelashes.

There are at least four possible solutions for droopy eyelids and downwards eyelashes:

1. If possible, move the camera or the mirrors to film the eyes from even further below.
2. Ask the participant to use an eyelash curler to turn the eyelashes upwards. This instrument should be desinfected between participants.
3. Is the participant tired? Ask him to return at another date and time that is more appropriate.

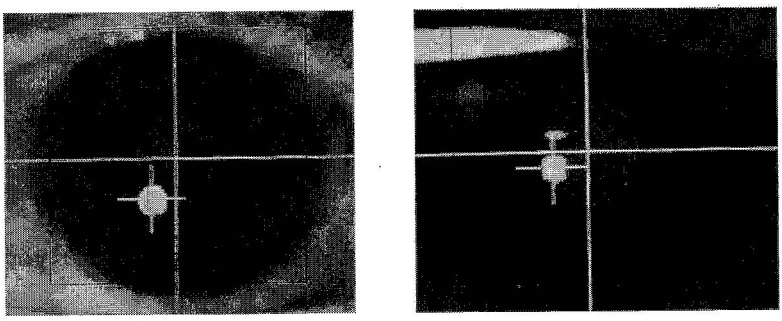


(a) Mild effects. The bottom five lines have a one-line offset, and the two last lines are smeared together into one thick data line.



(b) Strong effect. Note the offsets and optic artefacts in the bottom four lines, and the centre. The data in the upper paragraph are fairly good, since the eye is open while the participant is looking there.

Fig. 4.8 Example of the effect in reading data of droopy eyelids.



(a) No glasses

(b) With glasses

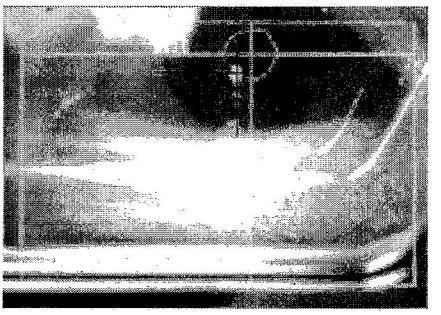
Fig. 4.9 Effect of glasses on one and the same participant. Note in (b) the extra reflection just above the corneal reflection, the darker image with lower contrast, and the large reflection in the upper left corner.

4. As a final resort, use some sticking plaster to fixate the participants upper eyelid. This works well for most elderly participants, but may not be comfortable for some participants.

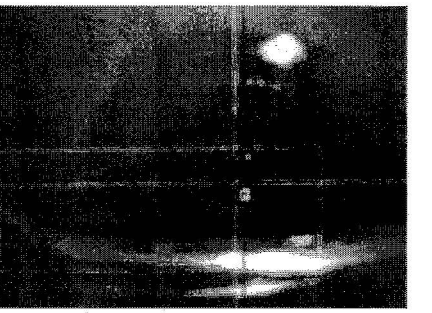
4.4.3 Shadows and infrared reflections in glasses

Glasses make eye tracking more difficult in several ways. First, glasses, or possibly the surface treatments of the glasses, may make the eye image darker, reducing the contrast between pupil and iris, which may decrease accuracy and precision in a dark-pupil system. Access to contrast and brightness settings in the eye camera is useful to remedy this. Second, the light from the corneal reflection can be reflected back to the eye and give a second, but fainter corneal reflection higher up in the eye image (compare the eye images in Figure 4.9). Third, the shadows from the brim may confuse the pupil detection in dark-pupil systems. All these three effects are seen in Figure 4.10.

Another major problem is infrared reflections in the glasses themselves. If the infrared reflection is on top of the pupil or near the corneal reflection in any of the gaze directions—ask your participant to look around—your calibration or your data recording will be jeopardized. This problem is particularly large if the participant wears old, scratched glasses, or glasses

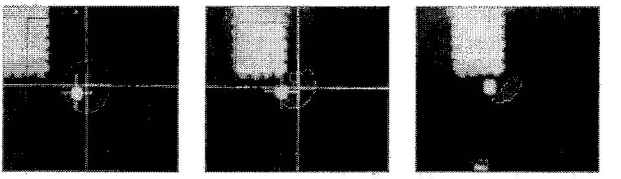


(a) Eye with shadow from glasses, and an infrared reflection high above the eye. Dark areas could confound the pupil detection but have been excluded from detection.



(b) Eye with multiple infrared reflections and refractions from glasses, both above and below the pupil. Were the eye to move, the pupil might be covered.

Fig. 4.10 Shadows and reflections in glasses.



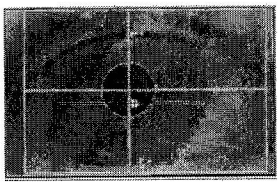
(a) Tracking works (b) Tracking with offset (c) Tracking lost

Fig. 4.11 Reflection in glasses covering more and more of pupil and finally making tracking impossible.

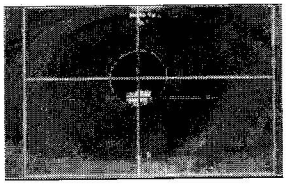
that have been treated to reflect sunlight, while glasses with anti-reflection treatment tend to elicit much fewer reflections. The solution is to move the eye camera angle so these reflections appear far from the pupil and corneal reflection (see Figure 4.11). It may take a while to find a view on the eye without interfering reflections, in particular with tower-mounted eye-trackers, but then these systems usually have good tracking throughout the experiment as the participant is fixed. For remote eye-trackers you are usually limited to changing the filming angle and then hoping that the participant does not switch to a disadvantageous position during the experiment.

Some people wear small glasses with thick designer frames. This may cause a dark shadow on part of the eye, which can interfere with both pupil and corneal reflection. Again the solution is to move the camera position and angle until an optimal viewing angle is found.

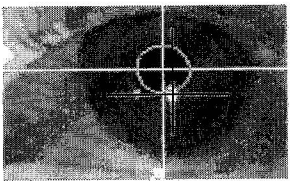
Some glasses are more difficult than others, in particular if they are so dark that the contrast in the eye video is too low, very small so the frame comes too close to the eye, or because they are very scratched and create many infrared reflections. A frame too close to the eye is problematic because it may occlude pupil or corneal reflection in certain eye positions, or the frame itself may be mistaken for a pupil or a corneal reflection if it is dark or reflective. If your eye-tracker allows you to change the luminance (and contrast) of the eye image, do that. Another solution, used by some eye-tracking researchers, is to have a set of their own eye-tracker-friendly glasses of different strengths that they lend to participants with problematic glasses.



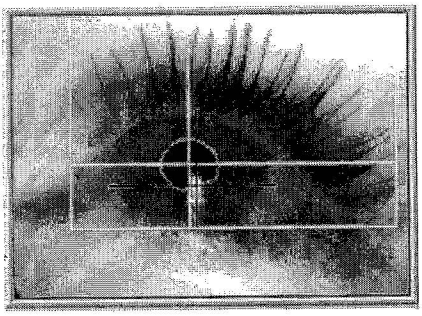
(a) No air bubbles at the point of the corneal reflection.



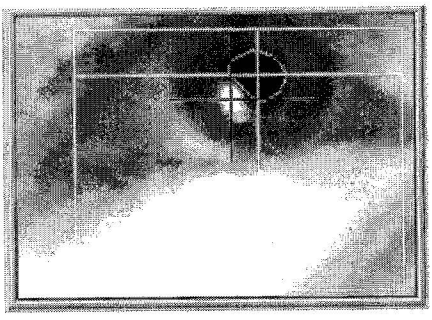
(b) The effect of air bubbles on the corneal reflection.



(c) Hard contact lens; no air bubbles, but if the border of the lens passes over the corneal reflection, a similar offset will occur.



(d) Focused camera. Two corneal reflections; cross-hair on the upper one.



(e) Unfocused camera. Several corneal reflections merged into one, so tracking is stable, but the pupil detection is deformed, giving a small offset.

Fig. 4.12 *Soft contact lenses* in (a), (b), (d), and (e): Reduce focus on camera to avoid multiple corneal reflections. The *hard contact lens* in (c) does not generate bubbles.

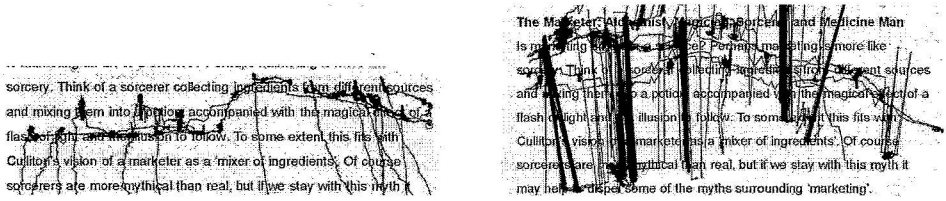
4.4.4 Bi-focal glasses

Bi-focal glasses introduce a border in the midst of the eye video that makes calibration difficult and recording close to impossible. Ask your participant to wear other glasses.

4.4.5 Contact lenses

Soft contact lenses cause only one problem in eye tracking, but it is a major problem, and it is not uncommon. For some people, because of a less than perfect fit between lens and eyeball, small air bubbles gather underneath the contact lens. When the infrared illumination is reflected in such a collection of small air bubbles, the light is split up into a number of reflections (Figures 4.12(b) and 4.12(d)). As the eye moves, the bubbles shift. Your eye-tracker will randomly select any of these reflections as the corneal reflection, and jump between them. This results in erroneous data samples and apparently very fast movements of the eye (Figure 4.13), known as optic artefacts. When this occurs it can be devastating for your data quality. You cannot always see the bubbles while setting up and calibrating a participant; they may appear in the midst of data collection, ruining your recording session.

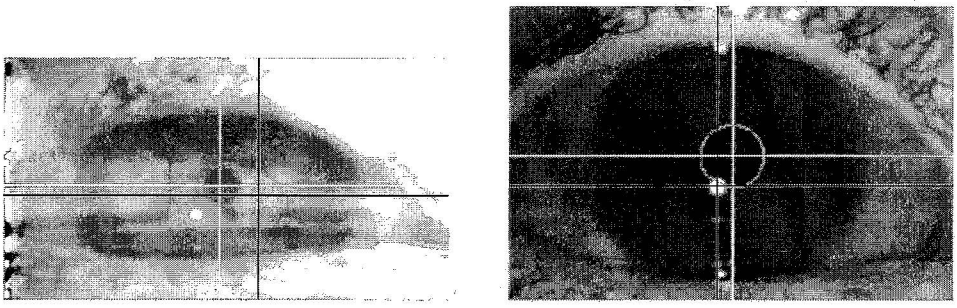
Fortunately the solution is mostly very easy. If your participant wears soft contact lenses, which you should ask them about or can even see in the eye image, always reduce the focus of the eye camera somewhat. This way the many small bubbles will merge into one larger corneal reflection. Your data will sometimes be slightly less accurate, because the larger corneal reflection pushes the pupil slightly to the side. Nevertheless, an unfocused recording is much, much better than if you had continued with full focus and a split corneal reflection.



(a) Participant reading one line of text.

(b) Participant reading two lines of text.

Fig. 4.13 Offsets and optic artefacts, due to contact lenses causing a split corneal reflection. The length of the false movement lines corresponds to the distance in the eye video between the two corneal reflections. As Figure 5.14 on page 163 shows, velocities in these artefactual movements is far beyond that of saccades.



(a) Full sunlight. Low contrast and high luminance; pupil is identified but corneal reflection is not.

(b) Indoors with only artificial illumination.

Fig. 4.14 Same participant wearing a head-mounted eye-tracker.

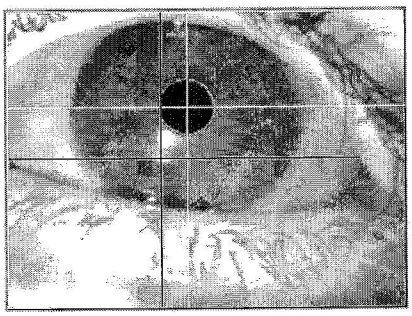
This will work in all cases, except the most extreme and rare ones, when the corneal reflection splits into a multitude of points that can cover up to a quarter of the iris.

Hard contact lenses are formed to fit the eyeball much tighter, and appear to gather no air bubbles, and data can be recorded at full eye camera focus. Hard contact lenses are very clearly seen in the eye camera image, as in Figure 4.12(c). Only if the border of the hard lens moves across the corneal reflection, which can occur often for some participants, will data quality suffer.

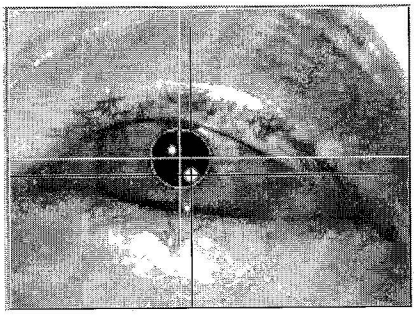
4.4.6 Direct sunlight and other infrared sources

Direct sunlight contains much infrared light, enough to outshine the infrared illumination many times over. The result is typically complete and immediate data loss (compare Figure 4.14). This can be particularly difficult during car driving, especially when the sun is close to the horizon, shining directly onto the eyeball. One of the authors was once in a car-driving project where we drove along the Scanian coast with a western sun shining through a sparse forest. The resulting rapid alternations between sun and shadow yielded equally rapid alterations between data capture and data loss. Such data, of course, are useless. In cars, one option is to attach an infrared filter film to the windows, but a more common option is to simply record on a cloudy day. The SmartEye systems that are specifically designed for use in cars select a part of the infrared frequency spectrum in which the sun has a lower effect on the eye-tracker camera, compared to the normally used infrared frequency.

Static systems can also be affected by sunlight, for instance through a window. Good shades or no windows at all are therefore recommended in an eye-tracking laboratory.



(a) A hot incandescent bulb gives a second reflection in the cornea, and the corneal reflection cross-hair has locked onto it.



(b) The fast-moving fourth Purkinje reflection in the upper left quarter of the pupil.

Fig. 4.15 Additional reflections.

A video-based motion capture system illuminates the scene to be measured with several infrared lamps, which could seriously compete with the infrared diode of the eye-tracker. We have on several occasions tested head-mounted systems with motion capture systems such as Qualisys. This combination has always worked. Obviously the shorter distance from the eye to the eye-tracker infrared diode more than compensates for its weaker luminousity.

In fact, some ceiling-mounted indoor lights can be more harming to eye tracking than motion trackers as measured by the intensity of the corneal reflection that they give rise to. In Figure 4.15(a), an incandescent light from the ceiling caused an additional corneal reflection, which is similar to the double reflection of the contact lenses, but only now it is permanent. If we were to calibrate on such an eye image, the danger is that some of the calibration points are calibrated with the correct reflection, and other on the false reflection, as in Figure 4.15(a). Terrible offsets result from such a calibration. Solutions include turning off the light, or using a lamp that does not emit infrared light,[14] and if that is not possible, use recording software settings to select properties such as reflection perimeter and distance to pupil centre to tell the eye-tracker which reflection to use.

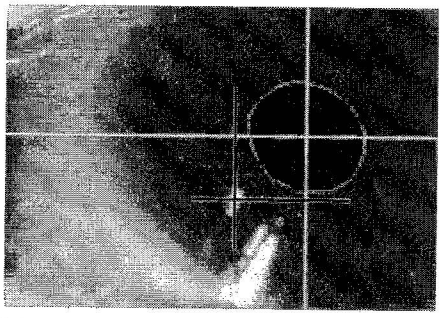
4.4.7 The fourth Purkinje reflection

You may sometimes see an additional reflection inside the pupil, as in Figure 4.15(b). It moves very quickly when the participant moves his eyes, but as soon as it reaches the border of the pupil, it disappears. It is produced at the back of the lens, and can therefore only be seen through the pupil. This reflection is used by dual-Purkinje eye-trackers, and is the reason they are so precise and have such a small tracking range. It may on rare occasions—for only a few samples—interfere with the corneal reflection and undermine data, but it is the weaker one and will not overly displace the cross-hair from the corneal reflection, as long as the eye camera is focused. If the camera is out of focus, this reflection may be much bigger and undermine the calculation of the pupil area.

4.4.8 Wet eyes due to tears or allergic reactions

When the participant's eye is wet, the corneal reflection splits up into several reflections, similar to the split corneal reflection in a contact lens. This may happen due to allergic reactions in the pollen season, or if you instruct the participant to try not to blink. The latter case makes

[14] You can check any lighting source for infrared by covering the eye-tracker's infrared source and pointing the camera at the light—if the system picks it up you know it emits infrared light.



(a) Allergic reaction resulting in a wet eye and multiple corneal reflections. The identification cross is on a false reflection.

(b) Infrared reflections in the retina giving a semi-bright pupil.

Fig. 4.16 Reflections in tears and retina.

them gaze until their eyes are dry and then the eye compensates with more tear fluid. The distance between reflections is often larger in this case, so the problem cannot be solved by decreasing eye camera focus. However, it mainly appears when gaze is in the far upper corners, as in Figure 4.16(a), and can mostly be solved during manual calibration with little or no data loss.

4.4.9 The retinal reflection (bright-pupil condition)

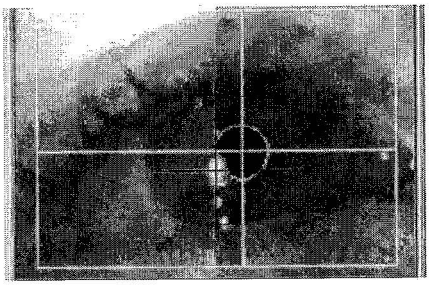
On rare occasions in dark-pupil systems, you may happen to have the infrared illumination almost co-axial with the line of sight, and then see the reflection in the retina. Unless you are using a bright-pupil system, you do not want this, because it makes pupil identification difficult or impossible, as in Figure 4.16(b). This condition can be solved by moving illumination and the mirror.

4.4.10 Mirror orientation and dirty mirrors

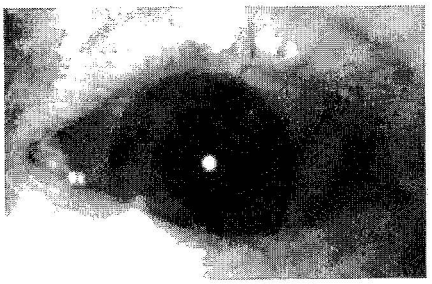
Wrong mirror orientation is a very uncommon problem, but one which may occur—it did in our laboratory. During the preparations for a large study (310 participants in six days using four tower-mounted systems), we were cleaning the mirrors of the systems, and happened to put one of them upside down. Now, the mirror that is part of many eye trackers, is covered on one side by a thin layer that reflects infrared light but lets through visual light. It is usually difficult to place the mirror upside down, but when it happens, by accident, in the hurry of the last preparations, the infrared light is spread into four corneal reflections along the vertical axis, as in Figure 4.17(a). The diagnosis was difficult, but once the problem was found, the solution was easy.

Specks on mirrors may give rise to weak reflections in the eye image that reduce the contrast and increase noise levels, as shown in Figure 4.17(b). The solution is simple: clean the mirror.

The examples above should be ample proof of the importance of the user being able to see the eye video and influence its quality. Hiding the eye video from the user, and relying on automatic set-up of the eye video, is a current trend in manufacturer development. It is an understandable ambition to attempt to make this part of data recording more easy, but if this automatization is not properly done, it can make eye camera set-up and recordings more difficult, give data of a poorer quality, and leave the user with fewer clues on how to improve data quality.



(a) Four corneal reflections along the vertical axis, as a result of an infrared-reflecting mirror turned upside down.



(b) To the left of the pupil, a weak reflection is visible, equal in size to the pupil, that reduces the contrast and will increase noise levels in recordings. Clean the mirror from fingerprints and specks.

Fig. 4.17 Mirror reflections and lowered contrast due to a dirty mirror.

The entire eye camera set-up and positioning of the participant should take no more than a minute or two once you start to gather experience.

4.5 Calibration

Individual calibration of each participant is necessary for a variety of reasons. For instance, the eyeball radius varies by up to 10% between grown-ups, and it may also have different shapes. Glasses alter the size of the eye in the eye video image. These variations change the geometrical values underpinning the calculation of gaze direction (Hammoud, 2008).

If you are using a head-mounted system out of the laboratory, see to it that the level of ambient infrared light is as low and stable as possible. If you calibrate in an area with a lot of infrared light, and then record in a darker area, the contrast between pupil and iris may change and result in poorer data, unless the camera compensates and the system dynamically resets the contrast thresholds. For maximal data quality, always calibrate in the same luminance conditions as you will have during data record, and minimize luminance variation in the trials (see Chapter 2).

4.5.1 Points

Calibration is typically made on a 2D area which has a number of predefined calibration points: 2, 5, 9, 13, and 16 are common numbers. Calibration points should cover the area where the relevant stimuli will be presented, whether it is a monitor, a scene video from a head-mounted system, or even just a part of a screen. Using the pupil and corneal reflection positions recorded in calibration points with known coordinates, the recording software can calculate a function to estimate any given location on the stimulus, given the extracted pupil and corneal reflection positions. Systems with just one or even no calibration points are being developed. By calibrating the system on just a few calibration targets, the system can then fit a function that allows it to interpolate between all intermediate positions and also extrapolate to positions outside the calibration area. It should be noted that accuracy is better within or close to the calibration targets and the stimuli should preferably be appearing within the area encompassed by the calibration points.

Points should be small and visually salient, so that participants gaze at them as exactly as possible during calibration. Any misalignment of gaze towards the actual calibration point

will introduce a corresponding offset in your data. Some laboratories use high-frequency Gabor patches that are only visible when participants look straight at them. Animated calibration targets seem to work better than static ones, as participants gaze at them as long as they are animated rather than looking around for the next target. Having the participant mouse-click on very small points rather than just looking at them may increase accuracy even further. For small children, it can be an advantage to exchange the standard point for objects that are fun or somewhat familiar to them: a yellow sun, a blue star, etc.

The position of points differs substantially between systems that give data files and those that only give gaze-overlaid video output, such as the head-mounted systems. For data files, the calibration points should be put on the stimulus surface (i.e. most commonly the computer monitor on which the experimental stimulus is displayed). For gaze-overlaid data, the calibration points should be in the coordinate system of the scene video camera. When calibrating the scene video of a head-mounted system, a laser pointer is commonly used to project an actual target overlapping the calibration coordinates determined from the scene camera. Alternatively, it is possible to position the scene camera at an exact position where known physical targets will coincide with the calibration targets of the system.

4.5.2 Geometry

There are a number of geometry settings that the calibration routine needs access to: the size (and position) of the calibration screen in pixels and/or millimetres, the precise position of the calibration points, and the distance from the eye to the monitor. These values can be set manually with some eye-tracking systems, but are automatically estimated or measured in some of the remote systems.

Monitor distance affects data quality in corners, and what really matters is how much of the visual field is covered. Unless the participant has difficulties accommodating, there is little benefit in presenting on a 24 inch screen at a position chosen so that is covers a visual angle equivalent to a 17 inch monitor at a shorter distance. You have to find the proper compromise between the position when the corneal reflection is lost in top corners, when the pupil is covered with eyelashes at the bottom corners, and between the distance and height of the monitor in the overall set-up. Even with a very good set-up, some participants' eyes will nevertheless cause problems in bottom and top corners.

4.5.3 The calibration procedure

Always test the quality of your eye video before calibrating. Let the participant look at the corners of the calibration area and watch whether the eye video looks good with a stable pupil and corneal reflection in all corners. Then you can calibrate.

The participant looks through all calibration points, and at each point the eye-tracker software samples a few hundred milliseconds of data. The participant in Figure 4.18 looks at nine different points. Notice how the spatial relation between pupil and corneal reflection differs in the different gaze directions. It is of vital importance for the quality of your data that the full and correct pupil and corneal reflections are selected at every single calibration point. An eye video set-up with the images in Figure 4.18 would give a successful calibration and good quality data.

The progress through the points, as well as acknowledgement that the sampled data are valid, can be done *manually* either by the operator or by the participant, or *automatically* by the recording software. The current trend among manufacturers is to provide automatic calibration as their default. Tests that the authors have carried out with close to 60 participants in each of the three conditions (i.e. operator controlled, participant controlled, auto-

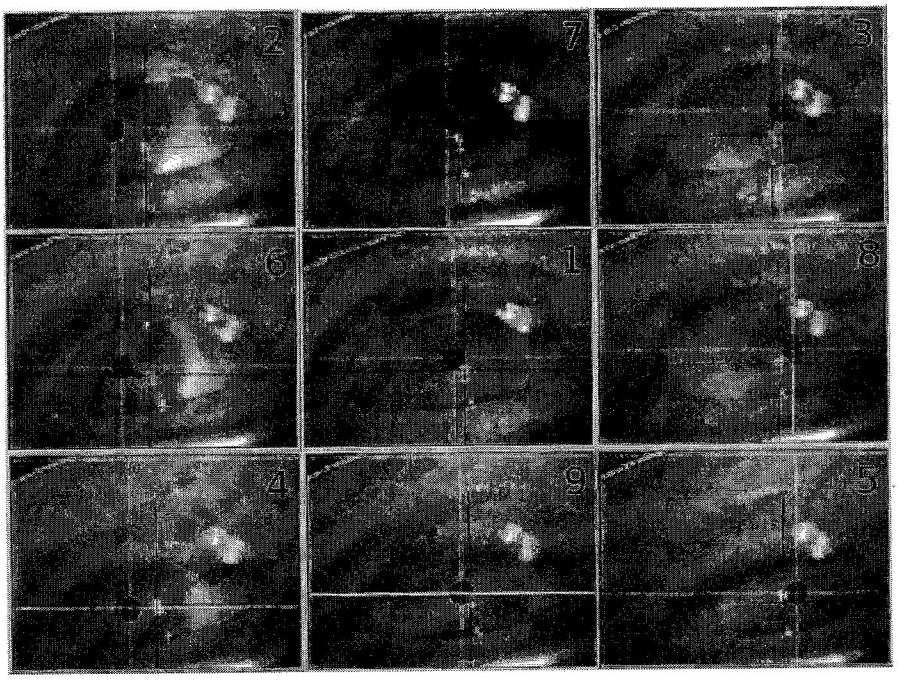


Fig. 4.18 Participant (with glasses) looking at the nine points of a calibration screen in the order given by the numbers 1 to 9. The bright reflections are always out of the way for successful detection of pupil and corneal reflections, although close in calibration point 3.

matic) showed that data quality (precision and accuracy) is superior for *participant-controlled* progress and acknowledgement of calibration points (Holmqvist, Nyström, & Andersson, 2011). In other words, participants know best themselves when they are looking at a point, and the researcher can hand over the calibration process with confidence.

Irrespective of calibration method, the participant must cooperate in looking at the calibration points, and they virtually always do. Exceptions may include children or animals who lose their interest in the calibration procedure after a few points. Patients who have difficulty keeping their eyes still, e.g. due to congenital nystagmus (common with albinos), are also difficult to calibrate.

You may sometimes miscalibrate participants simply because they happen to look in the wrong direction while you or the eye-tracker acknowledges one of the points. Sometimes, a participant tends to fixate the calibration points for a very short time, not knowing that you need a second or so to verify the image in the eye video and press the acknowledgement button. Other participants may make involuntary so-called square wave jerks during calibration (see Figure 5.30), giving an offset because the participant looked away slightly from the calibration point just at the moment that point was confirmed in the calibration procedure. To reduce the danger of such errors, always instruct your participants to keep their gaze fixed in the centre of each calibration point until it disappears, and watch out for participants from groups that are known to have a higher rate of square wave jerks (p. 407).

4.5.4 Corner point difficulties and solutions

Many systems allow you to watch the eye camera image while calibrating, so you can see that the eye is still, that it is looking in the right direction, and that corneal reflection and

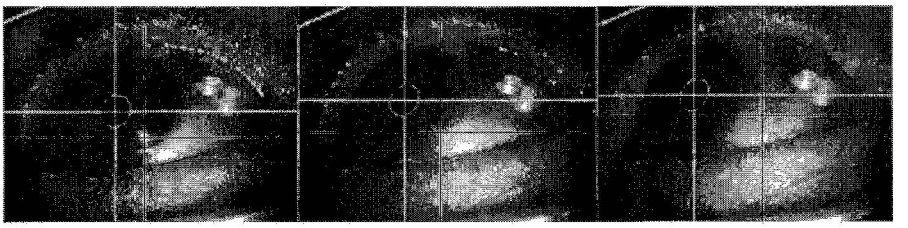


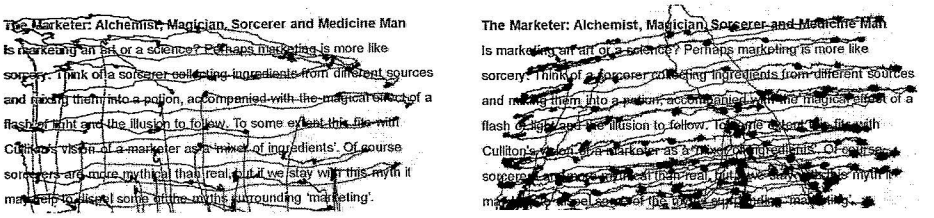
Fig. 4.19 More and more extreme gaze directions towards the upper left calibration corner. The first eye is OK, the second is dubious, and the third will give a considerable offset in data, as in Figure 4.20.

pupil are correctly identified, as in Figure 4.18. The most difficult calibration points are the corners. *Bottom corners* are problematic if your participant has downward eyelashes partially or completely covering the pupil, when looking down at the bottom corners. In that case, there are three things that you can do:

1. Ask the participant to open his/her eye. This will temporarily keep the eyelid or lashes from blocking the eye feature detection and lets you proceed with the calibration. But during recording, when your participant is fully occupied with the task and no longer thinks about holding the eye open, you will get optic artefacts, offsets, and data loss at that bottom corner. This may be a viable option if the problem only concerns the corners of the screen and the interesting parts of the stimuli are mainly located in the centre.
2. Acknowledge calibration of that point even if half the pupil is covered. This is gambling that your participant's pupil will also be half-covered during recording. Your data will have a smaller offset compared to the corners in case one, but the poorer precision and accuracy will remain, and will likely contain optic artefacts. If you are planning to do area of interest-based analysis this may be OK, but fixation and saccade measures will be affected.
3. Go back and set up the eye camera to film from further below, or ask the participant to curl the eyelashes. This takes more time, but it is the only guarantee of good data. With a remote, try tilting the camera or moving it closer or further from the participant to change the angle.

The *top corners* are usually easier, but they have a problem of their own, illustrated by the eye images in Figure 4.19 and the resulting data in Figure 4.20. When the corneal reflection moves across the border of the iris, it changes position and form. If you calibrate with a corneal reflection outside of the iris, then you will later have a considerable offset for data samples in that corner. There are of course large individual differences between participants as to the gaze direction at which this corneal reflection leaves the iris. For some participants with a narrow eye opening, the corneal reflection is instead covered by the lower eyelid. In any case, there are three things you can do to remedy this situation:

1. The quick solution is—if your recording software allows it—to move the calibration point further in, and ask your participant to look at the new position, hopefully thus moving the corneal reflection back into the iris. This solution will give precise and fairly accurate data when recording, especially if the task normally does not require the participants to look at extreme angles but rather in the centre of the screen.
2. You could also move your eye camera and/or infrared illumination so as to film the eye from further up. This takes more time, and you also run the risk of creating problems at the bottom corners if you change the camera angle too much.



(a) Upper right calibration point (b) Upper left calibration point

Fig. 4.20 Offsets—poor accuracy—resulting from calibrating participants whose corneal reflection had left the iris, as in Figure 4.19. In 4.20(a), the upper right calibration corner was problematic. In 4.20(b), the upper left calibration corner. These errors are easy to spot in a highly structured design and task, such as in reading, whereas in other tasks it will be much more difficult to identify these offsets.

3. The most proper solution is to position the stimulus (monitor) so that both top and bottom corners are OK for the large majority of participants. This you have to do when building the recording environment and setting up the geometry of your experiment. A greater distance from the participant to the monitor will lower the maximum visual angles required and alleviate these problem, but it will also lower the difference in angles between different areas of interest and lower the accuracy and precision.

4.5.5 Calibration validation

It is important to never take a calibration at face value. A participant may have shifted his eyes just as you or the system captured the eye as representative for that particular calibration marker. The simplest solution is, right after the calibration, to show an image containing the calibration markers and ask the participant to look at the markers while you verify them visually on the control computer.

Some systems provide a numeric accuracy value of the average deviation between markers and gaze position. This accuracy value should be reported when you submit your article or report, but in the past this has been unusual. High-end systems exhibit values around 0.2°, whereas a maximum average deviation of 0.5° should be demanded for most studies. An average deviation of 1.0° would be unacceptable for instance for reading research investigating preview benefits and word landing positions, and some remotes produce 1.5° or larger average offset. If you use this number, be sure to relate the reported visual degree value to your particular stimuli display: how much off can the data be before your analysis suffers? Recalibrate until the validation values is below your required accuracy. Only then start the data recording.

It is important to keep in mind the difference between the estimated accuracy and the real accuracy. Even if you perform a validation, the validation points will also suffer from a random or systematic error. A realistic goal would be to have the offsets for the calibration/validation targets as small and random, i.e. offset equally in all directions, as practically possible. For studies that require very high accuracy, there exist correction methods that can be used immediately subsequent to calibration. For instance, Santini, Redner, Iovin, and Rucci (2007) describe a method in which raw data samples are shown on top of the calibration points, and a joystick is used to alter the underlying transformation matrix so that data are right above the points.

Although precision is just as important to validate as accuracy, in practice precision validation is very uncommon. An exception is Santini et al. (2007).