Chapter 3

Practicalities of Eye-Tracking

3.1 Designing an Experiment

Designing and building an eye-tracking study can be a daunting task. In this chapter we will walk through the practicalities of planning, creating and running a study from beginning to end. The specifics of how to actually build an experiment vary according to the type of study we want to run and the eye-tracking system we are using. Our aim is not to provide a 'how-to' guide as such, but to talk about the key components and general issues that will be important to most studies. We will also give a brief introduction to the three systems that were discussed in Chapter 2: SR Research EyeLink (www.sr-research.com), Tobii (www.tobii.com) and SMI (www.smivision.com)[1] eye-trackers. As with most things, the best way to learn how to build an eye-tracking study on any of these systems is to dive in and try for ourselves. A good approach is to first look at an existing, similar study – if possible observing an experienced user as he/she builds an experiment – to see the 'dos' and 'don'ts', then to try editing and building a study of our own.

The first question we should ask is whether eye-tracking is even the best methodology to use. It is tempting to adopt new technologies even when a different approach might produce more useful data, so we should first consider the alternatives to decide whether this is the best option for our research purposes. As with any experiment in any field, the research question should inform the methodology and design, rather than the other way around. Reading extensively around our topic will show us how other researchers have chosen to study it, and this should help us to decide whether eye-tracking is a good choice. Assuming that this decision has been made (and that the answer was yes), this chapter addresses some of the things we will need to consider to get started. We will begin by describing the process of preparing stimuli and building an experiment, then discuss factors such as selecting suitable participants, setting up and calibrating the eye-tracker, and collecting and saving data.

First of all, we need to try to conceptualise our experiment as clearly as we can. Normally we start with a set of general questions that are not specific to eye-tracking, but which are important when designing any language study. Jiang (2012, Chapter 2) provides an introduction to the kinds of things we should consider in reaction time research, as do Colantoni, Steele and Escudero (2015, Chapter 3) in the context of investigating spoken language. These include identifying a research topic and question, preparing stimuli and finding suitable participants, all of which we will deal with here.

3.1.1 Choosing the Type of Experiment

Most eye-tracking studies will follow a fairly uniform structure. Participants will be shown a series of visual stimuli and the eye-tracker will record the eye-movement patterns for each

[1] As noted elsewhere, SMI has been purchased by Apple. It is unclear how their systems and software will be updated and supported in the future.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

person, along with any other user inputs such as key or button presses, mouse movements and any verbal responses made during the experiment. These could include the participant 'naming' something on the screen, or our study might be interested in using a 'think-aloud' approach, where the participant describes his/her thought processes during the experiment.

The first step is to decide on the type of stimuli we are interested in using and the 'task' we will be presenting to participants (that is, what exactly we will be asking them to do). Broadly, stimuli will be either 'static' or 'dynamic'. Static stimuli will be things such as text or still images, while dynamic stimuli will include moving images. In a typical text-based study, the task will be simply to read the text on screen. This might be made up of single sentences or of longer extracts, depending on our research question (see Chapter 4 for more on the questions we can address with text-based studies). Very often we will want to include comprehension questions, either after every item or at certain points in the study. These can be included to genuinely test how well the participant has understood the text, or they can just be there as a way of ensuring that people actually read the items properly – people tend to pay more attention when they know they will be tested on what they are reading. As we will see in Section 3.2.1, reading studies can generate a lot of different eye-tracking data that we can use to understand how the text is processed. This makes it important to understand what variables we are manipulating and what we will be measuring before we begin, which in turn will help us decide which of the various eye-tracking measures are appropriate.

Image-based studies can take a number of forms. In applied linguistics, it is unlikely that our experiment will just use images and nothing else; a more common scenario would be a study that uses multimodal input (see Chapter 5 for some examples). This might be in the form of a 'storybook' study, where we are interested in whether including images alongside a piece of text improves comprehension for children or second language learners. Alternatively, a reading-while-listening task would present each piece of text on the screen and play a pre-recorded spoken version at the same time. As with text-only studies, we might want to include regular comprehension questions as a way of checking either understanding or attention, and we might also want to include some kind of test after the main experiment. This can be integrated into what we present using the eye-tracking software, or in some cases it might be easier to present it as an offline paper and pencil test after the eye-tracking is completed.

Some studies will combine visual stimuli with an audio track either by using static images or video. The 'visual-world' paradigm is discussed in more detail in Chapter 5 and will involve either a set of images (as in the Chambers and Cooke (2009) example in Chapter 1) or a visual scene (as in the Altmann and Kamide (1999) example in Chapter 1). Some studies (e.g. Holsinger, 2013) use a text-based variant, presenting different words on the screen to see which one participants are more likely to look at in response to the audio stimulus. In any study combining visual and audio stimuli, the data will normally need to be 'time-locked' to a specific point in the audio. This means that we will be interested in what our participants look at when they hear a specific word and for a specific period of time immediately afterward (typically a few hundred milliseconds). We can also define different time windows (often 50 or 100 ms time windows starting at the onset of a critical word) to see how listening unfolds over time. We address this more in Chapters 5 and 7.

Dynamic stimuli will generally either involve the presentation of a video or require the participant to interact with the display in some way, for example by looking at a website,[2]

[2] Not all websites will be 'dynamic', since in some instances we might just ask participants to view a static webpage that does not change. In other cases, we might need participants to scroll up or down a page, click on

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

Table 3.1 Main stimulus types in applied linguistics eye-tracking studies.

Stimulus type	Example study	Other possible elements
Text only	Reading study	Comprehension questions, during or after main study.
Text + image	Illustrated storybook	Comprehension questions, during or after main study.
Text + audio	Reading while listening	Comprehension questions, during or after main study.
Image + audio	Visual-world paradigm	Questions about the scene and/or audio input.
Video	Learning new vocabulary from watching videos	Comprehension questions after the video; pre- and post-video vocabulary test.
Website	Surfing to find information	Think-aloud recording.
External software	Using a corpus tool	Think-aloud recording.

Note: In most studies we are also likely to want to collect some background information on our participants (demographic data such as gender and age), which can be done either as part of the study or separately. We are also likely to need to use this data to match participants on a number of variables, which is discussed in detail in Section 4.3.2.

taking an online test, or using a piece of software to achieve a specific task (such as taking part in a chat session or using a corpus tool to produce concordance lines). Video-based studies are similar to visual-world studies in that we are interested in what the participants look at during specific time windows, or when they hear certain things in the audio track. We might also be interested in how participants make use of features like subtitles (e.g. Bisson et al., 2014). Here the task is likely to simply involve participants watching a video. We might want to include some questions afterward either to test their comprehension or to test understanding of newly introduced words. Again, this could be included as part of the experiment (administered by the eye-tracking software) or conducted separately afterward. For stimuli such as websites or external software where we are interested in the interaction of the participant with the stimulus, we might need to consider what to measure and analyse based on broad questions or hypotheses.

In any task, what we ask participants to do will have a big influence on their behaviour. If the task is to take a reading test or use a corpus tool to produce concordance lines, the instructions should be clear in informing participants of this, and of any time limits or other restrictions. For potentially less directed tasks, such as looking at a website, a task should be clearly defined, for example, 'Find *x* piece of information as quickly as possible.' This will be important to avoid participants simply surfing aimlessly, and therefore producing data that is difficult to interpret. Such tasks may also involve a think-aloud protocol or retrospective recall. Participants can be asked to commentate on what they are doing during the task, or else be shown the recording of their eye-movements afterward and be asked to commentate on it. Audio recording of this commentary can be done concurrently in some of the eye-tracking systems, but in others it will be necessary to make the recording ourselves using a separate device. The broad types of experiment likely to come up in applied linguistics research are summarised in Table 3.1.

links, etc. For ease of reference here, we consider all websites as 'dynamic' stimuli, unless static screenshots are used as the stimuli – see Section 3.1.2 for more on this.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

3.1.2 Selecting and Controlling Stimulus Materials

The stimuli we use are a key part of any experiment, so it is vital that we spend ample time choosing, developing and refining an appropriate pool of items. This is where reading around the research topic can be of great use. It can be very helpful to see what stimuli other researchers have used, and this can be used to help develop our own items. All published research should be replicable, so there is no problem with using stimuli from an existing study, provided we reference these appropriately to acknowledge where they were taken from.

Text Stimuli

In most text-based studies, we will identify a linguistic feature of interest and create stimuli that manipulate this. The analysis will focus on reading patterns for a word, sequence of words or longer stretch of text. A long history of research in psychology has shown us that eye-movements are affected by a range of features at the lexical or single-word level, the syntactic or sentence level, and the discourse or paragraph level. This is discussed in detail in Chapter 4. Clearly, controlling every possible variable in our study is not going to be possible, but we should certainly aim to balance the major factors that could have an unintended effect on our data.

Broadly, this means we should try to match critical words on features like length, frequency and part of speech.[3] We should also aim to compare different words in the same or similar contexts. When we manipulate a variable of interest and hold everything else constant, we can be confident that any difference in reading patterns is due to our manipulation. For example, in the study looking at the effects of metaphorical language introduced in Chapter 1, we are specifically interested in the reading patterns for the critical word in each sentence (which are underlined),[4] and everything else is the same (Example 3.1).

Example 3.1 Comparison of reading for a literal word ('affect') and a metaphorical word ('infect'), keeping all other aspects of the sentence the same.

The violence had even begun to affect normally 'safe' areas.
The violence had even begun to infect normally 'safe' areas.

If we are using longer stretches of text – paragraphs, or full-page extracts – we need to consider the position of critical words in the passage itself. Words tend to be read more slowly at the end of a sentence or paragraph due to 'wrap-up' effects (Rayner, Raney and Pollatsek, 1995). Effects of 'spillover' are also often seen, where the processing of one word or sentence carries over to the next, for example when an ambiguous word causes a reader to move forward in search of information that might be helpful. We should make sure that our stimuli are designed so that reading patterns for critical words aren't 'contaminated' by such effects. This means that where possible we should avoid having critical words at either the start or end of sentences, or in the first or last sentence of a paragraph. In

[3] Assuming our focus is individual words. If we are interested in phrases or longer sequences, we should consider the factors that are specific to these – see Chapter 4 for an in-depth discussion of this.

[4] Regions of interest are underlined in our examples but in a real study would not be visible to participants. See Section 3.1.4 for more about adding ROIs to experiments.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

Example 3.2, 'even begun to' would be the pre-critical region, 'affect'/'infect' our critical word, and 'normally "safe" areas' the spillover region.

Example 3.2 Critical words ('affect'/'infect'), pre-critical regions ('even begun to') and spillover regions ('normally "safe"') are all areas that we might want to analyse.

The violence had even begun to affect normally 'safe' areas.
The violence had even begun to infect normally 'safe' areas.

The general layout of the text is also something we should consider. The font style and size, line spacing, line breaks and margins should all be set in such a way that will ensure clear presentation and accurate data collection. For most studies black text on a white background will help make the presentation as 'natural' as possible, but for some specific contexts the colour scheme may need to be altered (e.g. a study with dyslexic participants might use pastel colours for the background to help reduce contrast levels). A font such as Courier New is often chosen since all letters take up the same amount of horizontal space. For most studies a font size of around 14–18 pt will be appropriate. Participants will be reading from a distance of between 40 and 70 cm from the screen (this varies by system; see Section 3.4 on setting up an eye-tracker), so the text needs to be big enough to be read, but not so big that it appears unnatural on the screen. Some testing will help to find the right balance. For multi-line reading, we may also need to edit the text a little to ensure that critical words are not at the start or end of a line. Aside from the problem of sentence wrap-up effects, fixations are not always stable when readers move from one line to the next, and often there is a tendency to under- or overshoot the target (the first word of the next line). In other words, a participant's gaze may not always move perfectly to the first word of the next line, hence minor corrective saccades may be made to this word, which would influence reading times. Some careful editing of the text size and the position of line breaks should help to avoid this for our critical words.

Eye-trackers are generally less accurate towards the edges of the computer screen, so we should set appropriate margins to avoid any problems. For single-line reading, each stimulus item can be presented across the centre of the screen, with a good size indent before and after. For full-screen text we can set margins all around the screen to ensure that there is no data loss as a result of the participant exceeding the trackable range of the eye-tracker. Since eye-trackers are less accurate at measuring vertical than horizontal movement, the line spacing should also be considered. Double line spacing, or even triple line spacing, is recommended to ensure that fixations can be reliably attributed to one line or another. In Figure 3.1, the fixation in version A cannot reliably be attributed to any particular line and could be on 'violence' or 'affect', whereas in version B we can be much more confident that the fixation is on the critical word 'affect'.

Some of the issues of font size and line spacing are only really relevant when we are investigating reading at the word, phrase or sentence level. If we are interested in the general reading of a text then we may not need to control the layout so closely. For example, if we are simply interested in global measures of reading and which part of the text readers spend most time on (rather than word- or even sentence-level processing), then having smaller text with less line spacing would not be a problem. See Chapters 4 and 6 for more on the different types of text stimuli we might use in eye-tracking and some of the methodological considerations that will be important.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

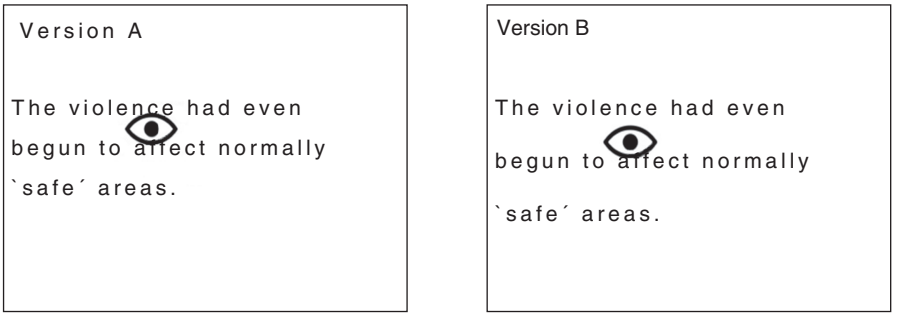


Figure 3.1 Example stimuli with single (left) and double (right) line spacing.

Image Stimuli

Any images we want to include in an eye-tracking study also need to be prepared carefully. This means that we should consider any potential confounding or distracting factors and do our best to minimise or balance these. Obvious considerations here are the size and quality or salience of the images we use (for a discussion of salience and other visual properties that should be considered when creating visual stimuli see Section 5.1). For example, we would not include some images that are simple black and white line drawings and some that are full colour photographs, since the photographs are more visually salient and would be likely to attract more attention. The choice of what type of image to include will probably be determined by our research question and the participants we will be using. For a storybook study looking at whether adding images to a piece of text helps with overall comprehension, we might choose different images for children than we would for adult second language learners. For children, colourful cartoon characters might be essential if they are going to take any interest in our study, but adults might find images like this distracting. There is also the practical question of where to get our images from. Clipart or royalty-free images can often be found online, but if we have specific requirements then we might need to consider whether we have the time and resources to either create these ourselves or commission someone else to do so. We should always consider the range of possible confounds (e.g. visual complexity, size, position on the screen) when using images to ensure that our materials are as balanced as possible. (See also the section 'Dynamic Stimuli' on preparing and balancing dynamic stimuli later in this chapter, and the discussion in Chapter 5.)

For images such as those used in a visual-world study, we need to consider the types of item we are interested in. Typically studies include one 'target' item (the picture we expect participants to look at), potentially a 'competitor' that is related in some way to the target, and then one or more distractors that are unrelated to the other items. In some cases, the other images may need to have specific properties, such as being semantically or phonologically related to the target item in some way (e.g. if the word 'beaker' is the target then 'beetle' would be a phonological competitor). As we have seen in Chapter 1, Chambers and Cooke (2009) investigated looking patterns to a target like 'poule' ('chicken' in English) and interlingual homophone competitors like 'pool'. However, unless this is the focus of our study, we need to ensure that our items do not contain any unintended cross-language overlap when working with language learners. Similarly, if we do not intend to study the effect of semantic or phonological overlap, we should make sure that our

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

'distractor' images are completely unrelated to the target images for all of the items we prepare.

For any images we want to present, the visual appearance should be balanced as far as possible. If we are combining objects in a visual-world study or visual scene, we need to make sure that the stimuli are all the same size and style, with no item appearing as more visually salient than any other. The different objects in the display should be spaced far enough apart that there is no confusion over what a participant is looking at. In the Chambers and Cooke (2009) stimuli, the images were placed at cardinal points (north, south, east and west) so fixations should be easy to identify. If we are combining text and images, we need to consider the position of the various elements and try to construct our stimuli in such a way that no one element is made more salient than any other. As with text-based stimuli, it is important to situate images in such a way that no data is lost at the edges or corners of the screen. Good tracker set-up will help to minimise this (see Section 3.4), but in general we should avoid any stimuli that will require the participant to fixate very close to the periphery of the monitor. This does not mean that we can't present images or videos full-screen, but we should try to ensure that 'critical' areas of the display are not close to the edge if we can help it.

Any images we want to include in our study can be prepared in advance using standard picture-editing software, then saved as single files for use in our experiment. Images should be of sufficient quality and resolution that participants will have no difficulty recognising what they are supposed to be, so we need to avoid using small or poor-quality pictures that will look grainy or blurred if we increase the size. Of course, while we can control many aspects of our stimuli when we are creating our own, if we want to use authentic materials such as a real illustrated storybook, we have less control. In these cases, selecting which aspects of the stimulus to analyse will be crucial, i.e. we may only want to select certain pages or extracts to compare.

Audio Stimuli

If our study also contains audio stimuli, some thought should be given to the recording of this. Will we record all items in the same voice? Are there any prosodic or phonetic cues that participants might subconsciously pick up on? For example, native speakers are able to discriminate short words ('cap') from the same syllables embedded in longer words ('captain') based on acoustic cues alone (Davis, Marslen-Wilson and Gaskell, 2002). Similarly, native speakers can reliably differentiate figurative and literal meanings of ambiguous idioms based on subtle prosodic contrasts (Van Lancker and Canter, 1981; Van Lancker Sidtis, 2003). We therefore need to consider how such cues could influence our results, and if so whether we need to modify the sound files in some way to remove these cues. Some studies record all stimuli, then cut the items up and re-edit them to minimise this possibility, although this may lead to items sounding slightly less natural. We should aim to avoid any systematic differences that might introduce additional variables into our study, such as having all of the critical stimuli recorded in a male voice and the other items in a female voice. Recordings should be clear and paced appropriately to ensure that participants will be able to hear and understand the stimuli as intended. See Section 5.2 for more on creating audio stimuli.

Importantly, when preparing a study that will combine audio prompts with eye-movement recording, the two elements will need to be accurately synchronised to allow us to

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

analyse the data. This may involve specific sound drivers, so we need to check the system requirements for both the computer set-up and the eye-tracker software we will be using. Some of the systems can combine visual and audio stimuli within the experiment (see Section 3.1.3 for more on this), otherwise we may need to prepare each stimulus item as a video where we add the audio recording over the static image. Video-editing software such as Windows Movie Maker can be used to create these so that the files can be added to the eye-tracker software when we come to build the experiment.

Dynamic Stimuli

For most studies using dynamic stimuli such as videos, webpages or an online test, it is likely that we are primarily interested in using authentic materials. This means that controlling our stimuli to the degree that we would with text or image stimuli is less of a priority, but there may still be areas that we need to consider (see Chapters 4, 5 and 6 for methodological considerations that might be relevant here).

Video stimuli can either be items that we have selected to suit the purposes of our study, or can be items that we specifically create with a purpose in mind. The first type is likely to be films or TV shows (or extracts from these). For example, we might want to show these stimuli to language learners or children as a way of testing their uptake of new vocabulary. In this case, it is likely that we will have selected something of an appropriate level and identified keywords in the video that we are interested in studying. We might therefore measure how much attention participants pay to areas of the display that correspond to these key vocabulary items, in which case we would need to ensure that they appear on screen for roughly the same amount of time and are roughly the same size, or that we have accounted for any such differences in our analysis (see Section 7.2). If subtitles are used, the language of the subtitles might be something we want to vary; for example, we could show learners of English a video with subtitles either in their first language or in the target language.

Alternatively, we might choose to create our own stimuli, for example by recording a lecture where new vocabulary items are introduced both verbally by the presenter and in writing on the screen. Controlling the position and duration of words would be an important consideration here, and all critical words should appear in comparable positions and for the same amount of time. Ultimately, we may need to decide whether using authentic material is more important than being able to control the stimuli as much as we would in other types of studies. The same may be true if we want to use other authentic materials, such as real texts (see Chapters 4 and 6 for more on this kind of study).

For other types of dynamic stimuli, how we control visual and linguistic aspects will depend on our research question and the materials we are using. We will need to decide whether creating something ourselves that can be carefully controlled is more important than using authentic material. For example, if we want to compare how participants process information on different websites, we will likely want to use real webpages. To do this we will need to be clear about the similarities and differences of the webpages in terms of the text, images, layout etc. For example, are we comparing a webpage with ten images to one with three images? If there are more looks to images in the former, this may be driven by the fact that there are simply more images. Using real websites and allowing participants to surf them freely is possible in some of the eye-tracking systems, but an alternative is to simply take a screenshot of the webpage we want to use and present this as a static image.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

In a similar vein, using corpus software to see how users interact with concordance lines might require us to give very specific instructions (e.g. for all participants to complete the same search), otherwise behaviour may be too diverse for us to draw any clear conclusions. In this case we might want to present participants with the software interface for a corpus search tool, then ask them to generate concordance lines for specific words or phrases. By doing this, we can ensure that their eye-movements and behaviour are in response to more or less the same stimuli. While there are good reasons to want to maintain the authenticity of our materials – for example if we are asking participants to take a standardised online language test – we should also be aware that how items are presented will have an important effect on eye-movements. This means that if we don't control the visual aspects of our stimuli – number and frequency of words, position and number of images, overall visual salience of critical aspects of the display, etc. – then the findings we can draw from the eye-tracking will be limited. Chapter 6 provides further discussion and examples of the kinds of studies we can undertake using authentic and dynamic stimuli.

Study Design and Counterbalancing

The type of design is an important choice in any study and will affect things like how we will present our items, and how we will analyse the data. In a lot of cases a 'factorial' design might be the best approach. This means that we will identify one or more 'factors' or 'conditions' and compare them using statistical tests (see Chapter 7 for more on working with your data). Examples of factors might be 'sentence type' (metaphorical or literal) in our earlier example of a text-based study, or word frequency (high vs low). Factors can also be related to our participants as opposed to our stimuli. Examples would be 'language status' (native speaker or language learner) in a study comparing two groups in a reading task, or 'reading ability' (high vs intermediate vs low) if we wanted to compare the reading speeds of school-age children according to the scores they had previously obtained on a standardised test. Often item factors and participant factors will be combined, so we might see a design described as two (sentence type) by two (language status), where we would show our metaphorical and literal sentences to a group of native speakers and a group of language learners to compare performance for each sentence type.

In some cases, a factorial experiment might not be as useful. This is likely for the dynamic stimuli we discussed previously, since carefully controlling the items to make two conditions might be much harder. It is still possible to create a factorial design, for example by showing videos with or without subtitles to see whether one leads to better comprehension and retention of newly introduced words. An alternative would be to take reading patterns for critical parts of the video and relate these to subsequent scores on a vocabulary test using either correlation analysis or linear regression. Baayen (2010) demonstrates how this kind of analysis is a good alternative for variables that have traditionally been analysed using a factorial approach, such as word frequency, but it would also be relevant for factors like language proficiency that can be graded.

A second important choice is whether our design will be within-subjects or between-subjects. In a within-subjects design, all participants will see all items, whereas in a between-subjects study, different participants will see different items. An example of a within-subjects design could be the video-based study mentioned in the previous paragraph. We could use two episodes of a cartoon and identify ten words (of comparable frequency) in each episode that we wanted to teach to children or language learners. We could show one episode with subtitles and the second without, then test whether the

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

presence of subtitles was beneficial when we tested the participants on a post-video vocabulary test. In an example like this, all participants would see both videos, so the design would be within-subjects.

Conversely, we might want to only show one video to each participant, but we might want to avoid showing the same video twice (once with subtitles, once without). In this case, we would show half of the participants the video with subtitles, and half without, then test the two groups to see if there was a difference in the uptake of vocabulary. In this case, the study would be between-subjects, since the two groups saw different versions of the stimuli. The same would be true in our example study of metaphorical language. We want to compare 'infect' and 'affect' in the same sentence contexts, but we would not want any one person to see both versions of the same sentence because repetition effects would be very likely to confound the results. To avoid this, we would aim to create multiple stimuli for each condition (multiple metaphorical and multiple literal sentences) and 'counterbalance' them over two separate presentation lists, as in Example 3.3.

Example 3.3 Counterbalancing of stimuli, so that participants do not see more than one version of the same item, but do see items from both conditions.

List A

The violence had even begun to affect normally 'safe' areas.
(Literal condition)
After all the rain the sunlight was bursting through the clouds.
(Metaphorical condition)

List B

The violence had even begun to infect normally 'safe' areas.
(Metaphorical condition)
After all the rain the sunlight was shining through the clouds.
(Literal condition)

Participants would be randomly assigned to one of the lists, therefore each person would see *either* the metaphorical *or* the literal version of each sentence. Crucially, because each sentence pair is matched so that the only manipulation is the critical word, the effect of metaphorical language (whether it requires greater processing effort than literal language) should be observable when we look at reading times/reading patterns across all trials.

The principle of counterbalancing is the same no matter how many conditions we have and no matter what stimuli we are using. In our between-subjects video example, we could easily introduce a second video and counterbalance the items, as in Figure 3.2. Participants would see *either* Version 1 *or* Version 2. Each person would therefore see one subtitled and one non-subtitled video, to enable the whole group to be tested on whether subtitles improved their vocabulary retention scores.

In some experiments we may want to include more than two conditions or factors. In a three-condition study there would be three versions of each stimulus item, so three presentation lists would be required and each participant would be randomly assigned to one list. Likewise, a four-condition study would require four lists, and so on. Items would be arranged in a 'Latin square' configuration so that each item appeared once and only once per condition across the required number of lists. Table 3.2 demonstrates how items

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

Table 3.2 Items counterbalanced in a Latin square arrangement. Each item appears once, and only once, per condition and per list. Each participant would see one list, and equal numbers of participants would see each list, i.e. 40 participants = 10 participants per list.

	Condition 1	Condition 2	Condition 3	Condition 4
List A	Item 1	Item 2	Item 3	Item 4
List B	Item 2	Item 1	Item 4	Item 3
List C	Item 3	Item 4	Item 1	Item 2
List D	Item 4	Item 3	Item 2	Item 1

Figure 3.2 Counterbalancing in a video study with subtitles. Participants would see either Version 1 (left) or Version 2 (right). All participants would see one video with subtitles and the other without.

would be distributed according to a Latin square design in a four-condition counterbalanced study. Each item appears once per condition and once per list.

Counterbalancing of this sort is only required when we have multiple conditions in our study and it is important that people are not exposed to more than one version of the same item. Within-subjects design of the type described previously are fine in many cases, and the decision is often not clear-cut. If we show different items to different participants, we run the risk that either specific items or specific subjects might be driving differences in our data. On the other hand, in a counterbalanced design the actual number of experimental items seen by each participant is smaller, hence statistical power is reduced. For a between-subjects study we will therefore generally need to prepare more items and plan to collect data from more participants than for a within-subjects design. The more conditions we intend to include, the more items we will need to prepare and the more participants we will need.

Counterbalancing can also be applied to other aspects of how we prepare our stimuli. If we are asking participants to do several tasks, we can counterbalance the order to make sure that this is not an unwanted effect (e.g. participants might always complete the first task more slowly than subsequent tasks). In the within-subjects version of our video example, all participants would see one episode of a cartoon with subtitles and a different episode without. In this case, we would want to ensure that half of the participants saw the subtitled episode first, and the other half saw the non-subtitled episode first. The same principle is true for any of the types of stimulus item discussed so far. For example, if we are asking participants to look at two different websites, we would counterbalance the order in which they were displayed; if we were presenting two extracts from a novel,

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

we would show extract 1 first to 50 per cent of the participants and extract 2 first to the other 50 per cent, etc.

For any image-based studies we should ensure that the position of critical images is counterbalanced across all trials. In the visual-world example from Chambers and Cooke (2009), objects appeared at the four cardinal points. In order to ensure that no effect of 'position' emerged, we should ensure that the critical target word appeared in each position an equal amount of times over the course of the study. That is, 25 per cent of the time it should be in the north position, 25 per cent in east, 25 per cent in south and 25 per cent in west.

How Many Stimuli?

The number of stimuli we need is largely determined by the nature and design of our experiment. We should aim to include an appropriate number of items to ensure that our analysis will have sufficient statistical power. Colantoni et al. (2015) suggest that a minimum of ten items per experimental condition should be the aim. What constitutes an item, however, will vary according to what we are presenting to participants.

In a text-based study presenting single sentences, or a visual-world study showing a picture plus audio stimulus, each stimulus will contain one 'item'. In other words, each critical word we are interested in will represent one data point in our analysis. For longer pieces of text – paragraphs or full-page extracts – we might include several 'items' in each page of text. If our stimulus was made up of a ten-page story, with two keywords and one image per page, this would provide us with twenty word- and ten image-based items. A video-based study might also contain multiple 'items', if the aim of our study is to test how well keywords that appear in the video are learned. For a subtitled video, the aim of the study will dictate what we are considering to be an 'item'. If we are interested in keywords, each word will represent one 'item'. If we are more concerned with general looking behaviour – how much attention readers pay to the subtitles in general – then we may want to define the occurrence of one 'set' of subtitles as one item, i.e. the period of time from when one set of subtitles appears on the screen to when it disappears is one 'item'. If the video lasts several minutes then we will have multiple individual 'items' to base our analysis on. If we are asking participants to view a series of webpages, each one might constitute one 'item', so we would want to include several different examples to ensure that we collect a usable amount of data. Example studies using a range of stimuli are discussed in more detail in Chapters 4, 5 and 6.

We therefore need to carefully consider the aims of our study, what we will be measuring and how we will analyse the data in order to decide precisely how many stimuli we will need to prepare. We also need to remember that for a between-subjects design, each factor in the experiment will double the number of items we need. For example, in our reading study comparing metaphorical and literal language, producing ten stimulus sentences would mean that each participant saw only five of each type once the items had been counterbalanced over two lists. We would therefore need to aim for a minimum of twenty sentence pairs to begin with to ensure that participants saw ten items per condition.

In many cases we will also need to include a number of filler items (non-experimental distractors) to mask the true purpose of our study. In both our reading and visual-world examples, filler trials would be included to ensure that participants were not able to spot any patterns in the stimuli (such as noticing that a lot of sentences contained metaphors). The number of fillers is normally equal to the number of experimental items in our study,

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

so a study with twenty experimental trials will likewise require twenty fillers. Fillers will not be needed in every kind of study. For example, showing participants an authentic video will not require fillers (we would not also show them other unrelated videos). If we are presenting text-plus-image stimuli or reading-while-listening studies, our stimuli are likely to be stories or longer texts, where including fillers may not be appropriate. In such cases we should ensure that our critical words are spaced far enough away from each other and evenly throughout the texts to avoid any participant spotting a pattern. In general, fillers are most likely to be needed in any study where we will present multiple stimuli in different conditions, such as a single sentence reading study, and where we think it is likely that a participant might spot a pattern if too many critical stimuli were presented in close proximity.

It is important to strike a balance when preparing our stimuli so that our study does not end up being overly long. If we include too many items, participant fatigue could adversely affect our results, but too few items will limit the analysis that we can perform. For studies where the length of each item is fixed (such as visual-world studies where the audio for each item is the same length for each participant, or video-based studies) we know in advance how long the study will take. For text-based studies or those where participants are given more freedom (such as using a website or external piece of software), participants will vary in how quickly they perform the task so the length of our study will vary. Pilot testing will be useful for establishing what an appropriate number of items to include is.

We should also consider who our participants are when planning the length of our study. Less skilled readers (language learners, children, and people with certain language impairments) will in general read more slowly than skilled readers. Similarly, young children will not sit through as many trials as adult participants, so we should plan to collect less data, and also to make the study as appealing and interesting as possible. For people with language impairments, reading page after page of text may be very challenging. In all cases, we need to carefully consider who our participants will be when we are designing our study. (See Section 3.3 for more on choosing participants.)

As a general rule of thumb, an eye-tracking study should be no longer than one hour. Most will be much shorter, and some may be necessarily longer. For example, if we are asking participants to watch a whole movie, the session will be determined by the length of the video we are showing. If we want participants to read a whole book, we might need to plan for participants to come in for several sessions spread over a number of days. We should always consider including breaks in our study if we will be asking participants to sit and attend to a computer screen for any longer than around thirty minutes.

Order of Presentation of Stimuli

The order of presentation can be an important variable and is one we should try to control. There are some contexts where the order of stimuli must be fixed, for example in the case of a text-based study where we want to display sequential extracts from a book or show a story over several pages. Videos are necessarily fixed in the order they will present stimuli (unless we design and edit them ourselves). However, in many cases having a fixed order will not be important, and it may be desirable to present our items in random order to minimise the chance of participants spotting any patterns. We should also consider that trial order is a variable that could potentially affect performance. In general participants will speed up over the course of an experiment and respond to later trials more quickly than earlier ones, so randomising stimuli will help to mitigate against this. Trial order is

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

often a covariate that is included in analysis to ensure that any such effects are minimised (see Section 7.2 for more on data analysis). Stimuli can be randomised in advance (using a random number generator), or pseudo-randomised, where we arrange the items into a specific order to ensure that items from the same condition do not appear next to each other. Alternatively, all three eye-tracking systems considered in Chapter 2 provide in-built methods of randomisation for stimuli. If our stimuli are divided into separate blocks, we may wish to either randomise or counterbalance the order of presentation for each block to further minimise any trial order effects. In cases where this is not possible (storybooks, movies), we could incorporate the order into our analysis. For example, if we are interested in keywords that appear during a movie, we could number them sequentially and include the order as a covariate in our analysis.

Other Aspects of the Study

Some other features of our study that will need to be considered in advance are the instructions, whether we will include any additional tasks (such as answering comprehension questions) and any additional information (demographic data, language background or proficiency information) we may require from our participants.

How we phrase the task instructions will be important. In other words, what will we actually ask the participants to do? In reading studies this may seem fairly obvious, but patterns of reading may differ according to whether the reading is silent or out loud (Rayner, 2009), and also according to whether the participant is reading for comprehension, skimming the text, reading with the intention of memorising, etc. If it is important that participants read each stimulus item only once, then we should tell them this before the experiment. For visual-world studies, participants may respond very differently according to the task and what they are told to do (Tanenhaus, 2007a). For example, the pattern of results will differ depending on whether it is an action-based task where a response is required (e.g. naming some aspect of the stimulus or finding a particular image as quickly as possible), or a passive task where the participant simply has to listen to a stimulus while viewing the image on the screen. Asking participants to perform a specific task may reduce variability in their behaviour and therefore make the results more comparable and reliable, so we should give careful thought to how we phrase our instructions.

In some studies – especially text-based studies – we may want to include regular comprehension questions as a way of ensuring that participants pay attention throughout. Typically, these are included after every few trials (e.g. one-third of items will have a question after them) and can be used simply to encourage attention throughout, or can be designed to actually probe understanding. Some studies may require participants to answer longer questions, for example reading studies combining eye-tracking with a more qualitative investigation. The three systems considered in the following section have different ways of including questions in an experiment. We may also want to record participants' verbalisations during the study as part of a think-aloud protocol (or afterward as part of a retrospective recall). Some of the systems we discuss in the next section allow us to build this in, otherwise we might have to think about a way to record participant verbalisations separately and match the recording to the eye-movement pattern once all data has been recorded.

Finally, in many cases it is useful to collect additional data about our participants, such as gender, age, etc. Depending on our participant population, we may need to think

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

carefully about this. Information about child participants may be better collected from parents or teachers. For language-impaired subjects we may need to collect information specific to their impairment, such as type of dyslexia. Some information, such as the location of a lesion in brain-damaged or stroke patients, may need to be collected in conjunction with a medical professional, and we should be mindful of the need for confidentiality.[5] For non-native speakers we will generally want to collect details about their first language and language learning background. Some studies may also find it useful to include some method for collecting additional information, such as a vocabulary test, a test of working memory capacity or a test of non-verbal intelligence. Some of these tests can be incorporated into the experiment for ease of collection (see the following section on what can be included in each of the eye-tracking systems), but in many cases it may be just as easy to prepare paper or electronic tasks in advance.

Other data that we may require can be collected in the same way; for example, if we need to collect familiarity ratings for our stimuli then participants can be asked to rate words or phrases for how well they know them. If this is required, we should ensure that it is done after the eye-tracking data collection so that we do not give participants any information in advance that might affect how they approach the main task. Finally, if our aim is to use our study to examine the relationship between eye-movements and linguistic performance (e.g. vocabulary knowledge, grammatical learning, comprehension), we need to think carefully about those post-task measures, and in some cases administering a pre-test might also be necessary. We could use an immediate post-test, or we might want to use a delayed post-test, in which case we would need to arrange for participants to come back for an additional session. Thinking through all possible aspects of our study is essential if we are to design and run productive, successful experiments.

3.1.3 Building and Testing the Experiment

Once we have decided on the type of study (text, image, video, etc.), and the nature and number of stimuli we will need, the next step is to build the experiment using the software provided with our eye-tracking system. In this section we provide a brief introduction to building an experiment in Experiment Builder (SR Research), Tobii Pro Studio (Tobii) and Experiment Center (SMI). Each of these software packages allows us to include a variety of tasks and stimulus types. Detailed guidance for all three systems is provided in their relevant user manual. In this chapter we refer to Experiment Builder version 1.10.1630 (SR Research, 2015b), Tobii Pro Studio version 3.4.5 (Tobii AB, 2016) and Experiment Center version 3.6 (SMI, 2016b). References to specific sections may vary slightly depending on the version of the manual being used.

Building a study from scratch can seem challenging at first, so a good place to start is to look at an existing experiment to get an idea of what the structure looks like. Both Experiment Builder and Experiment Center come with example files for text, image and video display studies (as well as some others). Sample experiments within Experiment Builder are helpfully annotated to explain what each element does, so working through the structure is a good way to get to grips with the process of creating a study. These sample experiments are also described in detail in the user manual. Demo files for Tobii Pro Studio are available from the Tobii website. In all three systems, example files can be

[5] Confidentiality is important for all populations, but when dealing with patients or medical data, there may be additional considerations and ethical procedures.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

re-saved and edited, so an easy way to start making our own studies is simply to find one that uses the same structure and add our own stimuli in. Projects should be saved with new filenames and if necessary unlocked before any changes can be made (the relevant documentation on setting up new projects in each system will provide more explanation of this). In all three systems, it is important not to add or remove files manually, e.g. in Windows Explorer, so any changes such as adding files should be made from within the software itself.

Building the Experiment and Adding Stimuli

Once our stimuli are prepared and we have created a new project (or re-saved an existing one), we can begin constructing the experiment itself. Almost all applied linguistics studies, regardless of the type of stimuli, will follow the same basic structure. There will be a set-up phase, which will include welcoming and seating the participant, providing initial instructions, then checking and calibrating the camera (see Section 3.4 for more details on setting up the eye-tracker). Next comes the experimental phase, where each stimulus item is presented. The basic workflow for the experimental phase is Stimulus → Trigger → Optional Follow-up, where the 'trigger' is any action that causes the current display to move on. Triggers are generally manual responses (such as a button or key press) or can be timed so that the display remains for a specified length of time. If we have included a follow-up task such as a comprehension question after each item, the trigger will cause this to be shown, then the participant will need to make a response before the next trial begins.

The experimental sequence will often involve a practice session where a small number of non-experimental items are shown to the participant to ensure that he/she understands the task. This is also useful as it will ensure that any teething issues occur during practice trials rather than during any trials where data are being collected. The practice session can be recorded if required, although this is generally not necessary as the data will not be included in our analysis. Following any practice trials, the main recording sequence will take place. This is the part of the experiment where we will show our stimuli to participants and record their eye-movements and any other responses they make (button presses, spoken responses, etc.).

Once all items have been displayed, we would normally include a final 'Thank you' screen to inform participants that the experiment has finished. We would then collect any other information we might require, such as demographic data, or (for language learners) information about language proficiency and use. In some of the systems this kind of data collection can be built into the experiment, but it might be more straightforward to simply collect this separately before or after the eye-tracking experiment using paper and pencil materials.

The basic structure of any eye-tracking study is summarised in Figure 3.3. The first part (welcome and set-up) and the final part (experiment end) will be the same for almost all studies. The experimental phase will vary according to the type of stimuli. Comprehension questions are often inserted between items in a reading study and may also be used in visual-world or other 'passive' viewing tasks, but are less common in 'action' tasks (where the participant needs to perform a specific action in response to a stimulus) or video studies.

We might choose to include a longer set of comprehension questions after each video as a way of measuring how well participants have understood what they have seen and heard. Depending on the type and number of stimuli we have and the amount of time it takes to

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

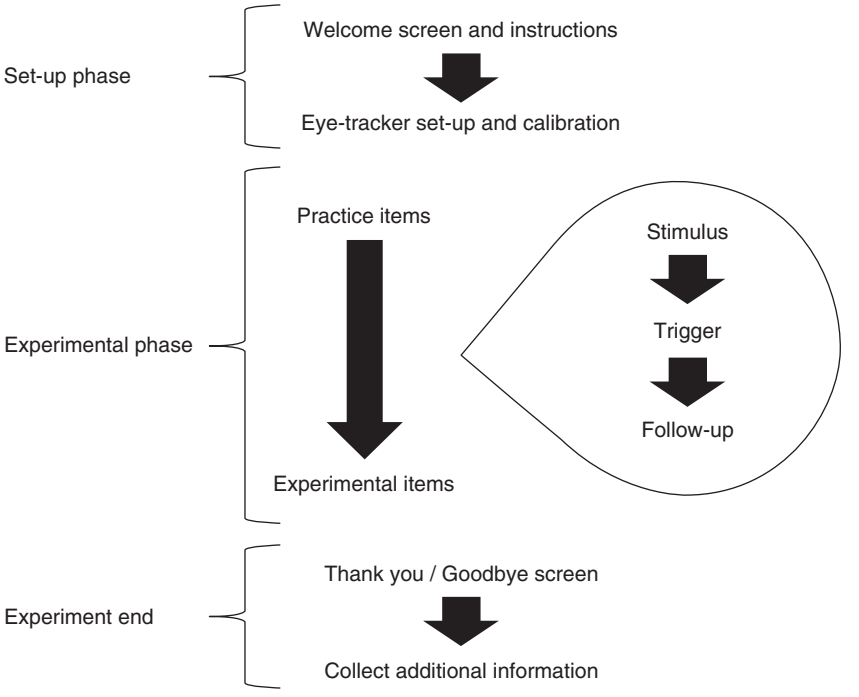


Figure 3.3 Structure of a typical eye-tracking experiment. The experiment will include an introduction and set-up phase, then the stimuli will be presented during the experimental phase. Each trial will consist of a Stimulus → Trigger → Optional follow-up, until all items have been seen. Usually we will include a set of practice items first, then show the experimental items. A final screen allows us to thank participants and inform them that the study has ended, then any additional information we may need can be collected.

present them, we might want to separate them into blocks and have a short break in between each block to prevent participants from becoming too fatigued (or bored) with a long sequence of items. This also allows us to build extra calibrations into the experiment to help ensure a high level of accuracy throughout. For example, if we have 150 sentences (experimental and filler items combined) in a text-based study, we might choose to present these in three separate blocks (and possibly counterbalance the order in which the blocks are shown to participants). If we are showing three extracts from a book, with each extract lasting ten pages, or if we are showing three five-minute videos, it would make sense for us to have a short pause and recalibration after each one. As with a lot of the considerations when building a study, a bit of trial and error is often required to figure out the best procedure.

Triggers

The triggers we use will be determined by the type of study we are running. In a text-based study we would normally want the text to stay on screen until the participant has finished reading, so we would include a manual trigger (button, key or mouse press) for participants to move on at their own pace. Since people read at different speeds, this is usually preferable to having the text on screen for a fixed period of time. In some studies we

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

might want to include a gaze trigger, whereby fixating a particular point (or word) on the screen causes the display to change or move on. This technique has been used extensively in the 'boundary paradigm' discussed in Section 4.5.2. Often more than one trigger may be used, for example combining a key press and a timer trigger. In this case the stimulus would remain on screen until a key is pressed, or if no key is pressed within a specified time limit, the trial will end. If this happens, it is usually an indication that something went wrong (e.g. loss of concentration on the part of the participant or a minor technical issue with that trial) so the individual trial should generally be removed from the analysis of the results.

For other types of studies, the timing of the display and the trigger required will vary according to the specific task. For a 'visual search' study, where participants must find a certain object within an image as quickly as possible, a manual trigger (e.g. a mouse press on an object) will be required in order to record how long each person took. Alternatively, a gaze-contingent trigger could be used, whereby fixating in a specific area for a pre-determined length of time will cause a display change or other action. In a storybook study combining text and images, again a manual trigger (e.g. press a button when you get to the end of the page) will be required to allow participants to read at their own pace and then move on when they are ready. Any image-based study that combines an audio stimulus (such as a visual-world task) will usually display each image for the duration of the audio plus a pre-determined amount of time afterward (e.g. one second after the offset of the audio) to allow us to examine how participants' eye-movements unfold over time.

Video stimuli will almost always be displayed for the duration of the video. No additional triggers will be required, although we could add one in case we wanted the option to abort the trial for any reason. For other dynamic stimuli, the type of task will again be important. Depending on what we are asking participants to do, we might want to present stimuli for a specified amount of time or until the user presses a specific key/button. See Chapter 6 for further information on how we might set up studies that use dynamic stimuli in this way.

The common trigger types are summarised in Table 3.3. This list only covers the trigger options that are available in each system by default. Additional functionality (e.g. using a button box in Tobii or SMI systems) can be achieved using certain plug-ins or by interfacing with additional software. These include plug-ins for commonly used programs such as E-Prime and Matlab and also open-source software such as EyeTrack or PsychoPy that can be of use in building experiments or adapting the basic capabilities of each system. We do not consider this here, but information on this kind of expanded functionality is provided on the manufacturer websites.

Building the Study

The three different systems we consider in this book (SR Research EyeLink, Tobii, SMI) have very different approaches to creating a study. In the following sections we provide a basic introduction to the process in each one. Working through the process of building an experiment is the best way to learn how to do it, and the user manuals provide much greater detail on all of the aspects we address here.

Experiment Builder (SR Research EyeLink)

Experiment Builder uses a drag-and-drop interface to design the basic layout of an experiment. The Graph Editor window is the main workspace and allows us to add in

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

Table 3.3 Common trigger types and their availability in the three eye-tracking systems discussed in this chapter.

Trigger type	Description	EyeLink	Tobii	SMI
Keyboard	Stimulus will display until a specified key is pressed on the keyboard.	Yes	Yes	Yes
Mouse	Stimulus is displayed until the mouse button is clicked, or until the mouse cursor enters a specified area of the display.	Yes	Yes (click only)	No
Button box	Stimulus is displayed until a specified button on a button response box is pressed.	Yes	No	No
Controller	Stimulus is displayed until a specified button on a hand-held controller is pressed.	Yes	No	No
Gaze	Stimulus will display until a fixation is made for a specified minimum length of time in a specific area of the display.	Yes	No	Yes
Voice	Stimulus is displayed until a pre-configured microphone detects an audio response (e.g. naming something on the screen).	Yes	No	No
Timer	Stimulus will display for a specified amount of time.	Yes	Yes	Yes

actions or triggers as required to create a flow chart. Separate 'blocks' (also referred to as 'sequences' in the user manual) are created containing each of the elements required to run our study. Figure 3.4 reproduces the useful schematic provided in the Experiment Builder user manual.

Within each block, dedicated 'nodes' represent each of the functions that we want to include. Each 'node' represents an action we want the software to perform, and is added to the Graph Editor workspace by dragging it into place. Nodes are connected to each other to create a sequence and generally consist of Actions (such as a Display Screen) and Triggers. The stimulus presentation sequence of an EyeLink experiment is shown in Figure 3.5, and represents the third 'Trial' and fourth 'Recording' blocks in Figure 3.4.

In this example, when each trial begins the software will automatically prepare the stimulus item (pre-load any graphics, reset triggers from the previous trial), then require a drift correct (a check on the participant's gaze position) to verify accuracy (see Section 3.5.2 for more on this). The recording sequence consists of a display screen (labelled here as 'STIMULUS ITEM' but by default called 'DISPLAY SCREEN' in Experiment Builder) showing the stimulus, which will remain until the participant responds by pressing any key (keyboard trigger) or until a specified length of time has elapsed (timer trigger). If a comprehension question is required, a second display screen could be added below the existing nodes within the Recording block. A second set of triggers would then be required, with the keyboard trigger specified to allow only certain responses (e.g. the 'y' and 'n' keys to allow participants to answer a yes/no question).

Any additional actions are added into the structure in the same way. For example, adding in the procedure to set up and calibrate the eye-tracker simply requires an EL CAMERA SETUP node to be added at the appropriate point (as part of the second block – 'Block' – in Figure 3.4). This set-up node should be added in at the beginning of

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

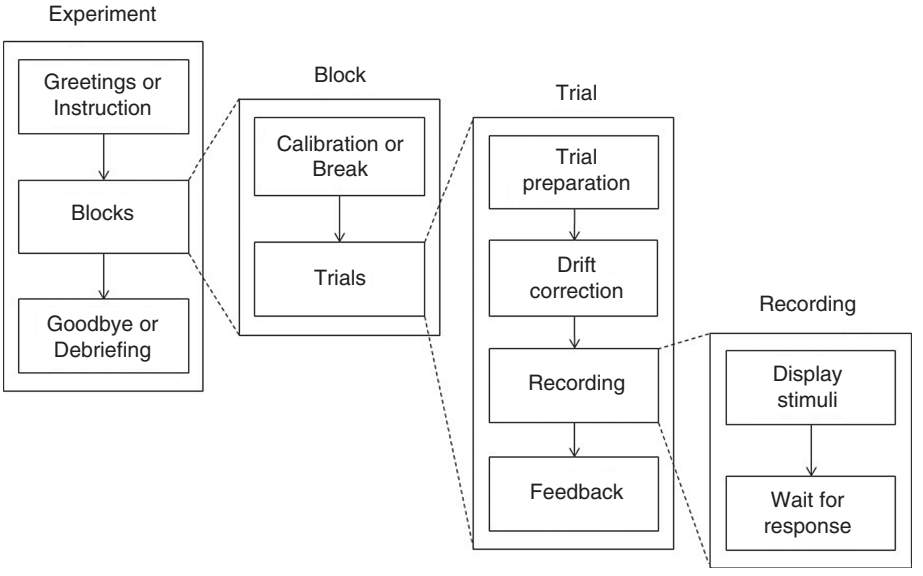


Figure 3.4 Hierarchical nested structure in an EyeLink experiment. Each 'block' is self-contained and forms part of the block above it in the structure. This allows an experimental block to be repeated as many times as required, i.e. the 'Trial' block would be repeated until all stimuli have been shown. Adapted from *SR Research Experiment Builder User Manual, Version 1.10.1630*, p. 36 (SR Research, 2015b).

any experiment we build, and by inserting it at this point rather than in the 'Experiment' block, a calibration procedure will be automatically included before each block of trials. This means that if we split our stimuli into separate blocks over the course of an experiment, each one will automatically include a set-up and calibration before the next set of stimuli are shown. All nodes and actions are detailed in the Experiment Builder manual, which provides a thorough explanation of the full functionality of the software.

Adding Stimuli In Experiment Builder we can easily add text, image, audio and video stimuli to our study. Once we have built the basic structure for our trial (as in Figure 3.5), we double-click the DISPLAY SCREEN node to open the 'Screen Builder' resource. Here, we add the required stimuli by selecting the corresponding icon, then edit the specific properties in the left-hand 'Properties' menu. If we want to include image, audio or video files in our study, these need to be added to Experiment Builder's library in advance to allow us to select them from within the project.

Text stimuli are added using either the Text Resource tool (for single-line reading) or Mutliline Text Resource tool (for multi-line reading). The text editor allows us to type in our text and set the properties such as font type, size and position on the screen. Image-based studies are created by using the Image Resource tool, selecting the required file from the library and then setting the properties such as width, height and location on the screen. We can combine text and image resources on the same screen if this is required, for example in a storybook study comparing the effect of pictures on reading patterns. Alternatively, text (with or without images) can be prepared and saved as an image file in

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

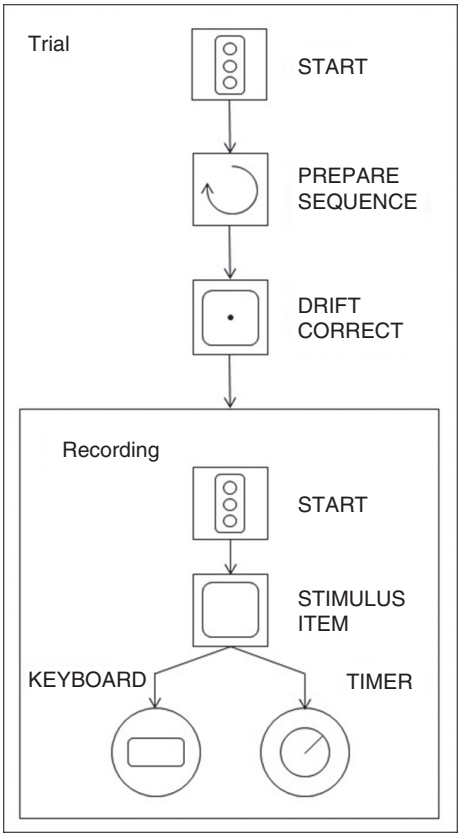


Figure 3.5 Sample layout for Block 3 – 'Trial' – and Block 4 – 'Recording' – within the overall EyeLink structure.

advance. For any study requiring an audio track to be played concurrently with either text (e.g. a reading-while-listening study) or images (e.g. a visual-world study), the 'Synchronize Audio' tickbox should be checked within the Properties list for the DISPLAY SCREEN node. This will enable us to select an audio file from the Experiment Builder library to play alongside the visual stimulus.

Video stimuli are added using the Video Resource tool. Supported formats are AVI and XVID. To ensure good video timing, it is recommended that video files are converted to the required format (XVID or VFW) using the Split Avi application provided with the Experiment Builder software package. Since this separates the video and audio streams, both must be added to the library. The 'Synchronize Audio' tickbox will need to be enabled for any video resource to allow us to choose the requisite audio stream to play alongside it. For any experiment requiring visual and auditory stimuli to be accurately synchronised (including a video-based study), specific drivers should be activated and used (detailed in Section 7.9.13, Playing Sound, in the Experiment Builder manual).

Experiment Builder does not directly support inclusion of some of the more dynamic stimuli discussed so far, such as webpages or external software applications. We can use webpages as stimuli by taking screenshots and adding them as image files. Alternatively, SR Research provide a Screen Recorder tool that allows us to record eye-movements

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

during other activities, such as surfing the web or using other pieces of software. More information on these is available via the SR Research support forum. For studies requiring a 'think-aloud' recording, Experiment Builder does support the recording of a participant's voice while viewing a stimulus. This may be useful if we want participants to describe an image they are viewing (e.g. a screenshot of a webpage).

If we don't have many stimuli (e.g. in a video-based study where we only want to show one or two videos), we can add them directly into the structure. In the majority of cases, this is probably not how we would add items. Most applied linguistics studies will require multiple stimuli, and adding 150 individual stimulus sentences in this way would be time-consuming and inefficient. The solution in Experiment Builder is to add a 'data source', which operates like a spreadsheet and which can be populated manually or by importing data from a text file. To do this we first need to create a prototypical trial (like the one in Figure 3.5), then link this to the data source containing our individual stimulus items. This is done at the level of the 'Trial' block in Figure 3.4 and is described in detail in Section 9 of the Experiment Builder manual.

Within the data source, each row represents one trial. If required, multiple items can be included in each trial by adding extra columns, and as many columns can be added as we need to specify additional variables, item groupings, etc. For example, having a column that specifies the 'Condition' for each trial (e.g. Experimental vs Control vs Filler) will make it easy to remove the fillers and compare the conditions of interest when we come to analyse our results. Text-based stimuli can be typed into the data source directly. If we want to add comprehension questions, we just need to add a column and type in the question that corresponds to each stimulus item, as in Figure 3.6.

Image, audio and video stimuli can be added to the data source by specifying the filename of an item that we have added to the library. Additional properties (e.g. if different images need to appear in different locations) can be specified by adding columns as needed. By default, Experiment Builder will present items in the order specified in the data source. If required, items can be randomised or pseudo-randomised in advance in the input file. Alternatively, Experiment Builder provides methods for internal and external randomisation, blocking and counterbalancing of stimuli. Section 9.6 of the Experiment Builder manual provides more information on how to randomise data sources.

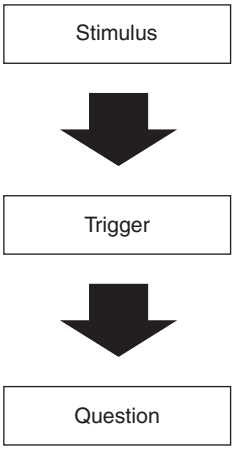
The Experiment Builder user manual includes a worked example (Section 14, Creating EyeLink Experiments: The First Example) and a very useful project checklist to ensure that the relevant steps have been completed for a successful experiment. We can use this, alongside the Test Mode available in Experiment Builder, to troubleshoot our experiment prior to collecting data. EyeLink experiments need to be deployed as a stand-alone .exe file (or .app for Mac users) before they can be used with participants, and once this has been done no changes can be made. If we do need to make any changes after this point, we would do this by amending our original Experiment Builder file and re-deploying the experiment.

Tobii Pro Studio (Tobii)

Tobii Pro Studio uses a hierarchical structure organised on three levels. The topmost level (Project) contains one or more Tests, with each Test defining a sequence of stimulus items. Within a Test, when a Recording is performed the eye-movement data is recorded and associated with the relevant participant and stimulus information. Section 3 of the Tobii Pro Studio user manual explains this organisation in more detail.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

Item	Condition	Stimulus	Question
1	Literal	The violence had even begun to affect normally 'safe' areas.	Was the city very peaceful?
2	Metaphorical	After all the rain the sunlight was bursting through the clouds.	Had the weather been pleasant?
3	Filler	The men had been chasing the dogs for several hours now.	Were the dogs in a cage?



The columns above correspond to 'Stimulus' and 'Question' DISPLAY SCREEN nodes that we would include in our prototypical trial. We now need to tell Experiment Builder which information to display by 'referencing' the required column in the data source.

To do this, we select the text resource that has been added within the DISPLAY SCREEN node, then in the 'Text' field of the 'Properties' list double-click to openthe Attribute Editor. Within the Node Selection structure, we select Trial > DataSource > then the column name that corresponds to what we want to display (Stimulus or Question).

An explanation and example of this is provided in Section 10.1, Using References, in the Experiment Builder manual.

Figure 3.6 Sample data source in Experiment Builder (top panel). Text stimuli can be typed in directly and will appear according to the parameters we set in the 'Text Resource' that we added to our prototypical trial. Other columns such as 'Item' and 'Condition' are useful for sorting our data once it has been collected and is ready for analysis. Once we have added a data source we need to 'reference' the required columns at the appropriate point in our experimental structure (bottom panel).

Creating a Test in Tobii Pro Studio involves dragging the required media element into place on a presentation timeline, then editing its properties. The simple drag-and-drop interface means that items can be reordered easily until all stimuli and any filler items have been included. Tobii Pro Studio supports a wide range of media elements, including instruction screens, images, movies, websites, screen recordings, questionnaires and PDF documents. Other options for external video recordings (where video from another source such as a webcam would be recorded) can be used, but these require external equipment. Figure 3.7 shows a sample layout in Tobii Pro Studio.

A text-based study can be constructed in Tobii Pro Studio in a number of ways. One is to use the Instruction element, which enables us to type in and format our own text. The example in Figure 3.7 uses an Instruction element both for the instructions and Text 1. If

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

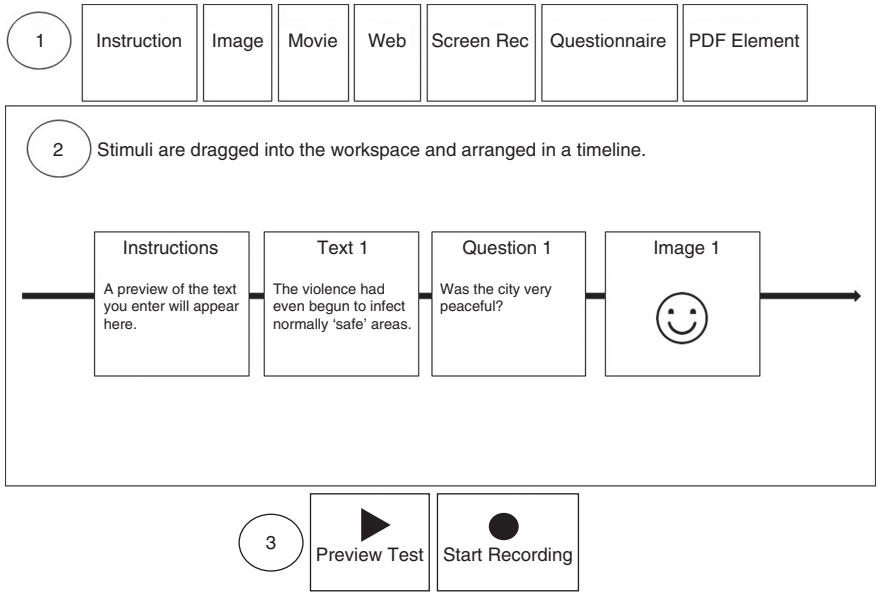


Figure 3.7 Schematic of a sample layout in Tobii Pro Studio. Stimulus items are added by dragging an icon along the top of the display (1) onto the workspace (2). This Test contains an instruction screen, a text stimulus, a comprehension question and an image. The order of elements can be changed by dragging to the required point on the timeline. Properties of each element are edited by double-clicking to open the media set-up dialogue box. At the bottom of the screen, the Preview Test button can be used to preview the experiment, and the Start Recording button will initiate a recording (3). NB: diagram is not to scale and some non-relevant icons are omitted.

we do this, we need to make sure that the 'Enable for Visualisations' option is ticked otherwise gaze data will not be recorded. Each of our stimuli (and each page if we are using multi-page texts) will require a separate element, so this might be impractical for studies with large numbers of stimuli. Alternatively, we can create images or PDFs for each of our items in advance and add these. For a text-based study where we want participants to read an extract over several pages, creating a multi-page PDF may be a good way to achieve this. If comprehension questions are required, these can be added using the Questionnaire element (multiple-choice questions only – no text input functionality is possible). A typical text-based study might therefore consist of Instruction element (stimulus) followed by a Questionnaire element (comprehension question), repeated as many times as required until all of our stimuli have been presented. The trigger for the stimulus element (to move the display on to the next element in the sequence) would be for the participant to press any key, then he/she would use the mouse to answer a multiple-choice comprehension question.

To add image or video stimuli, we drag the required icon into place on the timeline. Double-clicking the element will allow us to open the element's properties, from where we can browse to and add the required file, then edit other properties for the stimulus presentation. There is no dedicated audio stimulus media element, so this means that if we want to combine audio and visual stimuli we will need to prepare these in advance. For example, to create a reading-while-listening stimulus, we could create an image of our text, then use this as the background while we record the audio over the top using any

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

video-editing software. The resulting video file could then be added to the timeline. The same principle could be used for a visual-world study, where we would prepare our images in advance, then add the prepared audio (edited in advance) over the still image to create a video file.

Tobii Pro Studio supports a range of dynamic options, so adding in web stimuli (allowing the participant to browse the web freely while his/her eye-movements are recorded) or a screen recorder (to enable us to ask participants to use another piece of software) is also just a case of dragging the required element into place and specifying the URL or identifying the location of the software that should be launched.

Various properties can be defined for each of the stimuli by double-clicking the element on the timeline. These include how each element will end (keyboard, mouse press or timed trigger), as well as things such as the size of images and the appearance of text. Tobii Pro Studio provides either an in-built counterbalancing option to vary the order of stimuli randomly across participants, or allows us to define and specify different Presentation Sequences if specific presentation orders of the test are required. Section 4.1 of the Tobii Pro Studio user manual provides in-depth information on how to add and edit the different media elements, and how to include counterbalancing or Presentation Sequences in our experiment.

The Preview Test option in Tobii Pro Studio allows us to review our stimuli and amend as required. Data can be collected from within Tobii Pro Studio as soon as all stimuli have been added and we are happy that the experiment is complete.

Experiment Center (SMI)

The basic functionality of Experiment Center provides a simple way to create a range of studies. Additional modules can be purchased for specific purposes. For example, the Reading Analysis module supports the creation and analysis of text-based studies, and the Observation module enables the recording of concurrent or retrospective think-aloud data from participants. The Experiment Center manual provides more information on the expanded functionality that can be added.

The Application Window interface in Experiment Center provides an easy way to build our study. Rather than having to create our own structure, we add each stimulus item individually to a master list. Items can be reordered by dragging them into the required position in the list. Figure 3.8 shows a schematic of a typical layout in Experiment Center.

Experiment Center supports a wide range of stimulus types. All stimuli are added by selecting the required icon from the toolbar, which will create a new item in the stimulus list. The properties that can be edited for each stimulus are the type (text, image, video, etc.) and the name of each item; whether it should have a timed or manual trigger; whether it should be sized to fit the screen; whether eye-movement data should be recorded or not (for some items such as instructions we may not require data to be recorded); what task it should be assigned to (so that we can create multiple blocks or different versions of our experiment); and what randomisation group it belongs to (to enable us to randomise our stimuli in various configurations). A mix of different stimulus types can be included in one list, so we can easily create a study asking participants to do different things.

Text-based studies can be created either by using the Text element, or by preparing our stimuli in advance as image or PDF files. Selecting a Text element will open a text editor, so each item can be typed in and the font, size, position, etc. can be specified. A Question element can be added, either as a way of asking comprehension questions, or as a way of

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

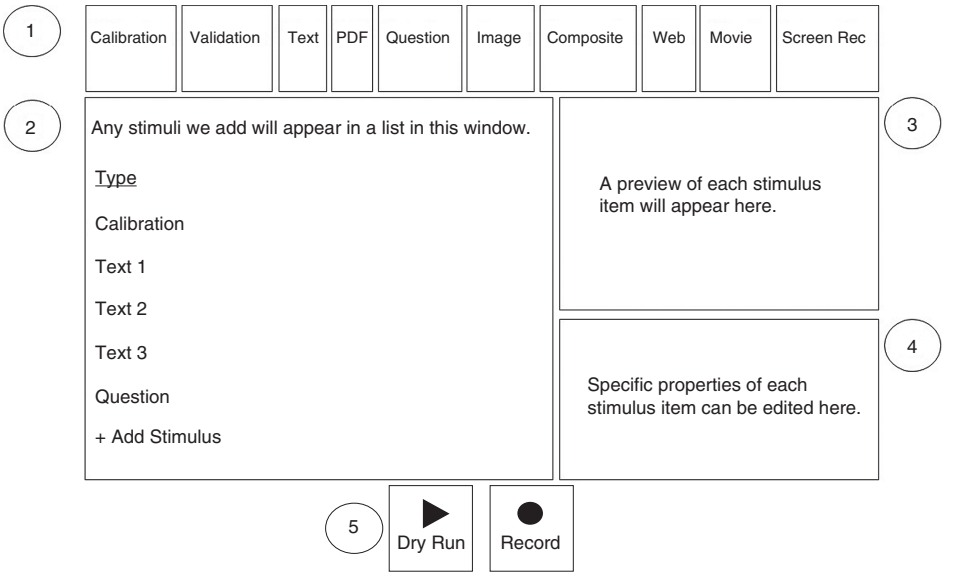


Figure 3.8 Schematic of the Application Window layout in Experiment Center. Stimuli are added using the icons along the top of the screen (1) and appear in a list in the left-hand panel (2). A preview of each item is provided in the top-right panel (3) and specific properties can be edited in the bottom-right panel (4). At the bottom of the screen, the Dry Run button can be used to preview the experiment, and the Record button will initiate a recording (5). NB: diagram is not to scale and some non-relevant icons are omitted.

collecting background data from participants. The answers can then be used to filter our participants during subsequent data analysis.

Image and video stimuli are added by selecting the required icon and then selecting a pre-prepared file on our computer. When adding images, if multiple files are selected each will be added as a separate item in the stimulus list. If we want to combine visual and audio stimuli (in reading-while-listening or visual-world studies), audio files can be easily added to most stimulus types as one of the fields available in the lower right-hand panel of the Application Window. Once we select the audio file that goes with the text or image, playback will begin once the stimulus is displayed. If a video stimulus is added it will be automatically encoded into an optimised format for playback. The Composite icon allows us to create a combined text/image/video stimulus, for example by uploading an image and then superimposing text over the top. If we have not prepared our items in advance, this would be an alternative way to create combined text and image stimuli (for storybook studies) or stimuli with multiple images (for visual-world studies), or to add keywords to a video.

If we have multiple stimuli, rather than adding them one by one we can upload a Stimulus List. Items can be prepared in a spreadsheet and uploaded together, which will be useful for a study involving a long list of items. Text, images and videos can be added like this, provided the required files are saved in the same folder as the spreadsheet. The spreadsheet must be prepared in a specific way, and Part VI Section 3 of the Experiment Center user manual provides details of how to do this for each stimulus type.

Once added, stimuli can be assigned to a specific Task. This provides an easy way for us to present different stimulus lists to different participants: all items can be added to the

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

overall project, then assigned to different Tasks to allow us to show our different counterbalanced lists (i.e. Task 1 would be items from List A and Task 2 would be items from List B). Both the order and duration of stimuli can be randomised from within Experiment Center using the 'Randomization Groups' option. If 'Task' and 'Randomization Groups' are left blank, items will be displayed in the order they appear in the stimulus list.

A calibration sequence is automatically included at the start of any experiment built in Experiment Center. If a validation is also required, we will need to add this in using the appropriate icon on the Application Window toolbar. Additional calibrations and validations can be inserted at any point in the stimulus list if required. For longer studies, regular recalibrations are recommended as this will help to maintain a high level of accuracy. (See Section 3.4.2 for more on calibrating and validating experiments.) The Dry Run mode in Experiment Center can be used to preview our experimental sequence and troubleshoot the experiment as we build it. Once we have finished adding stimuli, the project can be used to record data immediately.

3.1.4 Defining Regions of Interest

Regions of interest (ROIs; also called areas of interest or AOIs) are the parts of the display that we are interested in measuring and analysing. In text-based studies this is likely to be a word, sequence of words or part of a sentence, but might also be larger areas, for example if we are interested in time spent looking at the whole text versus time spent looking at images in a study combining visual inputs. In image-based experiments ROIs can be a specific area of the display, such as the individual images for any given trial displayed in a visual-world study. For studies that combine text and images – such as a storybook experiment – we might be interested in both specific words and images as our ROIs. For any experiment where the stimulus will unfold over time (multimodal studies where an audio stimulus is combined with text or images, or video-based studies), our ROIs will be defined as specific areas of the display at specific time points or during fixed time windows. In most studies, having a clear idea of what we want to analyse is a key part of creating balanced and controlled stimuli. It follows that our ROIs will generally be decided in advance for each stimulus item. However, for certain types of dynamic stimuli – for example writing or translation tasks where we are interested in the output produced by participants – we may not know what our ROIs will be in advance since the output will be generated by each specific individual. Examples like these are discussed in Chapter 6.

ROIs in Text-Based Studies

For text-based studies, it is particularly important to define our ROIs carefully to ensure that we obtain the most appropriate data, but also to ensure that we aren't swamped with data that we don't need. Both EyeLink and SMI systems can identify each word in a text-based study as a separate ROI. (In SMI this is only the case if the Reading Analysis module is licensed. If so, it will identify separate ROIs for each paragraph, sentence, word and character in our study.) While this can be very useful in some studies (e.g. as a way of measuring global reading behaviour such as the average fixation duration or saccade length throughout the text), it can be less useful for the kind of manipulation we are often interested in. For example, in our text-based study looking at literal and metaphorical sentences, we are primarily interested in the reading times for our critical word: 'affect'

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

versus 'infect'. We might also be interested in other areas of the sentence, such as the pre-critical and spillover regions, giving us three ROIs. If we allow the eye-tracking software to automatically treat each word as an ROI, analysis of these ROIs becomes less straightforward, as in Example 3.4.

Example 3.4 We can either identify our own ROIs in advance (a) or allow the software to identify each word as a separate ROI (b).

(a) The violence had even begun to infect normally 'safe' areas.
(b) The violence had even begun to infect normally 'safe' areas.

If we treat each word as a separate ROI, we have three potential issues. The first is that our pre-critical and spillover regions are segmented into separate words. This means that in order to analyse the whole region we would need to apply some kind of formula in the exported data to combine these into the ROIs we want. In comparison, in (a) we can easily identify the fixations on the pre-critical and spillover regions. The second problem is that by segmenting every word, we are producing much more data than we actually require. In this example, we have three data points in (a) but ten in (b). If we scale this up to include all of our stimuli (ten metaphorical sentences, ten literal sentences, twenty fillers) and participants (ten participants per stimulus list = twenty overall), we would end up with 8000 data points in the auto-segmented version, compared to 2400 in the version where we added our own ROIs. It is not hard to imagine how this would quickly become unmanageable in a study that included longer texts, such as paragraph or whole-page reading, where there might be perhaps 150 words per page.

The third reason to identify our ROIs carefully is that they are important for identifying regression patterns in the text. We will talk more about regressions in the following section (Section 3.2), but one thing we might be interested in is whether reading 'infect' in our text caused participants to return to the earlier word 'violence' to ensure that it had been correctly recognised and understood (see Example 3.5).

Example 3.5 Sequential numbering of ROIs is important for certain measures like regressions. In this example 'violence' would be ROI 1, and 'infect' ROI 2.

The violence had even begun to infect normally 'safe' areas.

Eye-tracking software calculates a 'regression' as a fixation on a previous ROI once the eye gaze has entered a later ROI (e.g. in Example 3.5, where 'infect' comes later than 'violence'). This means that for us to be able to analyse specifically whether 'infect' prompted more regressions to 'violence' than 'affect', these need to be clearly and sequentially identified as our ROIs. If we segmented each word into an individual ROI, the software would consider any regression to 'violence' from a word further on in the sentence, and any regression from 'infect' to a word earlier in the sentence. We might therefore not be able to tell whether it was specifically once 'infect' had been read that readers went back to read 'violence' again.

The ROIs we define will be determined by the nature of our research question. For text-based studies, whether we define single words as ROIs, or whether we want to identify phrases, clauses or even longer stretches will depend on what we are investigating. Longer stretches of text might be of interest in studies concerned with syntactic or

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

discourse-level processes. We consider reading measures in more detail in Section 3.2 and in Chapter 4.

ROIs with Visual and Audio Input, Video and Other Types of Stimuli

For image-based studies we are likely to identify specific parts of the display as our ROIs. As we have noted, most applied linguistics studies are likely to combine a visual display with some kind of audio input, for example in the visual-world paradigm. In such studies, we are interested not just in where the participant looks, but also how eye-movements change over time. Similarly, in a video, we are likely to be interested in specific aspects of the display that are on the screen at certain times of the video.

In a visual-world study, we would identify separate ROIs for each of the images that make up the display. In the Chambers and Cooke (2009) example, this means a separate ROI for the four images at the north, south, east and west positions. We will therefore be able to see the number and duration of fixations for each ROI throughout the trial. However, as we discuss in more detail in Chapter 5, visual-world studies are generally concerned with looking patterns at specific points in time. For example, Holsinger (2013) defined two windows of 400 ms each: an 'Early Window' (starting 180 ms after the onset of the critical word and ending at 580 ms) and a 'Late Window' (starting at 580 ms and extending to 980 ms). If we want to do the same, we have two ways to do this. The first is simply to draw each ROI and collect data for the whole trial, then afterward manually divide our results into any fixations that occurred between 180 and 580 ms, and any that occurred between 580 and 980 ms. Alternatively, we can set our ROIs to only be 'active' during certain points in the trial. As we discuss in the following sections, the three pieces of software we have considered allow us to define start and end points for our ROIs. This means that we could draw ROIs around the images on our display, then tell the eye-tracking software when they should start and stop collecting data. When we export our results, fixations would automatically be divided up into ROIs by location on the screen and by the time window in which they occurred.

For videos we are also likely to want to define ROIs that are only 'active' at specific points (i.e. they will only include fixations during a certain time window). For example, Bisson et al. (2014) looked at how much attention participants paid to subtitles while watching a DVD. They defined two broad areas corresponding to the subtitle region (roughly the bottom third of the screen) and the image region (roughly the top two-thirds of the screen). They then used the subtitle timing information to determine whether fixations occurred in the subtitle region while a set of subtitles was actually present on the screen. Data was only analysed for those time windows when subtitles were visible, allowing Bisson et al. to determine the relative amount of attention paid to the subtitles in each of their conditions (for a full discussion of this study and the conditions they manipulated, see Section 5.4.1). As with the visual-world example, we could conduct a study like this and set multiple ROIs to be 'active' during specific time windows (e.g. the start and end point of each set of subtitles). We could go further and identify specific words in the audio input that we were interested in, then define our ROIs to collect data only when those words or phrases occurred during the video, for example if we wanted to look at how well participants learned unknown vocabulary from watching the video and reading the subtitles. In this case we could define ROIs for the specific words, then relate overall reading times for these to subsequent scores on a vocabulary test. Figure 3.9 demonstrates the different ways in which we could analyse looking patterns to subtitles.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

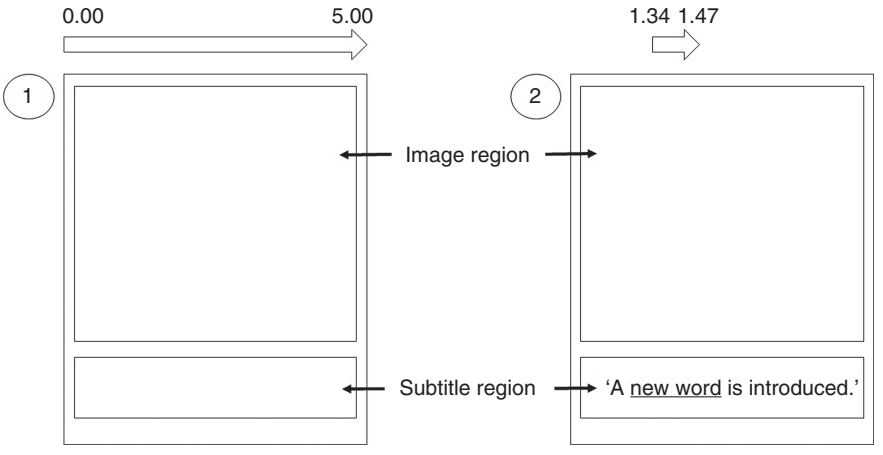


Figure 3.9 Example ROIs on a video with subtitles. In (1), two broad ROIs are defined for the image and subtitle regions. Fixation data will be collected throughout the video. The timing information can be used to filter our results into periods when subtitles are on the screen (as in Bisson et al., 2014). In (2), a time-locked ROI is defined that will only be 'active' from 1:34 to 1:47 – the period in the video when a specific set of subtitles appears on screen. Multiple ROIs with start and end points would be defined for each set of subtitles that we wanted to analyse. If we wanted to go further and look at a specific 'new word', we could create an ROI around its specific location in the subtitles to see how long participants spent reading it.

In videos and other dynamic stimuli, we may need to set ROIs that follow moving elements of the display. This is easiest if the motion is fairly smooth and linear, and we probably wouldn't want to try to define a dynamic ROI that tracked a complicated set of movements over an extended period of time. Although it is unlikely that we would need to track moving stimuli like this in the types of studies we consider here, we do briefly address the ways in which moving ROIs in each eye-tracking system can be created in the following section.

For some types of stimuli it may not be possible to define ROIs in advance. This is likely to be the case for software applications or webpages where the participants have a relatively free choice of what they can look at, in which case we do not have a predefined image to draw our ROIs onto before data has been collected. A good example would be someone using a chat window, where the text that appears will vary from person to person (see Chapter 6 for further discussion of this). In such instances, it would be necessary to wait until after data has been collected before we could identify and add ROIs. We should still have a broad idea of what we are looking for in advance, as this will help to ensure that our study design is robust and not simply a 'fishing expedition'.

Creating ROIs in the Three Eye-Tracking Systems

Here we provide a brief overview of how to create ROIs in the three systems we have considered so far. For clarity, we use the terminology adopted by each system: interest areas (IAs) in Experiment Builder/SR Research EyeLink systems, and areas of interest (AOIs) in both Tobii and SMI systems.

In Experiment Builder, IAs can be created either in advance when building the study, or following data collection using the dedicated SR Research analysis software Data

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

Viewer. For text-based experiments, each word can be identified as a separate IA by enabling the Runtime Segmentation option in Experiment Builder, or once the data has been collected by auto-segmenting trials in Data Viewer. As we pointed out, this may not be the optimal approach depending on the focus of our research, and may be most beneficial for obtaining global measures, or for longer reading studies with authentic texts where we are not able to create and control specific manipulations. In many cases it will be better to identify specific words and phrases in advance and draw in the IAs ourselves (see earlier discussion about Example 3.4). Other types of stimulus can also be 'auto-segmented', where Data Viewer will automatically create a grid of IAs for us, for example over an image. Often this will not be specific enough for our purposes, so we would again want to define our own IAs for analysis.

Adding IAs in advance can only really be done where we have very few stimuli, or for studies where the position of our critical elements will be the same for all items. For text-based studies where words will vary in their length and position in the sentence, this will be tricky. One solution is therefore to deploy our experiment and run through the final version ourselves to collect 'live' data. We can then draw in all the IAs we require in Data Viewer by selecting one of the Interest Area Shape icons and using the mouse to drag a box around the region we want to define as the IA. These Interest Area files (text files stating the name of each IA and the *x* and *y* coordinates on the screen) can be saved within the Experiment Builder project directory in the Library → Interest Area Set folder. We can then add a column to our data source to specify which IAs should go with which stimulus items. Re-deploying the experiment will mean that these IAs are automatically created in Data Viewer for any new data that we collect. This process can be a little confusing, but the Help menu within Data Viewer provides more information and an example. An Interest Area file can also be created whereby the ID, shape and location (*x* and *y* coordinates) of all of the IAs we require for our study are specified in a text file. This can be added to Experiment Builder prior to deploying our experiment, or to Data Viewer to add IAs once data have been collected.

Both time-locked and moving IAs can be created, and Data Viewer refers to these as 'dynamic' IAs. To create dynamic IAs, we need to deploy and run through our study once, then in Data Viewer we draw the required IAs onto each of our stimuli. Drawing static and dynamic IAs is done in the same way, but for dynamic IAs the 'Dynamic IA' tickbox needs to be checked. Dynamic IAs by default require us to specify a start and end time, which defines when it will be active during each trial. Moving IAs that track part of a video stimulus can be created in a number of ways, but one option is to use the Mouse Creation tool. In this, we create an initial IA, then during playback of the video we can click on the display at consecutive time points to follow the element we are interested in and create any subsequent IAs that are required in different locations. Slowing the playback speed or pausing the video at short intervals will help here, especially if we are following something that moves quickly or non-linearly in the display. Section 8.3, Interest Areas of the Experiment Builder user manual and Section 5.9 of the Data Viewer user manual (version 2.4.1, SR Research, 2015a) provide further information on how to create and add IAs in EyeLink experiments.

In Tobii Pro Studio, AOIs can be drawn manually for each stimulus item before or after data has been collected using the dedicated Areas of Interest tab. For 'non-interactive' elements (Instructions, Images, Video), once we have added our stimuli to the timeline they will be available to add AOIs, and any that we draw in will be listed on the left-hand pane. For 'interactive' media elements (Websites, Screen Recordings

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

and PDFs),[6] the element will not be available for us to define AOIs until after data has been recorded for one participant. Each AOI we add can be set to be 'active' or 'inactive' at specific time points. We simply select the AOI we want to edit, then use the time indicator slider underneath the display to set the time points where we want it to start and stop. AOIs can be set to track an aspect of the display by selecting the starting time point and drawing in the required AOI. If we then move to a later time point and move (by dragging) or resize the AOI, Tobii Pro Studio will automatically calculate the intervening frames to ensure that the display is tracked in a continuous fashion. (NB: this interpolation is linear so will assume that the movement is in a straight line and at a constant speed. Any exceptions to this will need to be corrected manually on a frame-by-frame basis.) AOIs can also be 'grouped' so that data from a larger area can be considered, e.g. to combine text that is located on different parts of the screen for an overall analysis. Defining and editing AOIs is dealt with in more detail in Section 8 of the Tobii Pro Studio user manual.

In SMI systems, AOIs are created after data has been collected using the AOI Editor in the analysis software BeGaze. Once we have added AOIs to our stimuli, these will apply for all subsequently collected data. It is therefore possible to collect data from one participant then add in AOIs, and these will be applied to any participants we test from that point on. For text-based studies, the Reading Analysis module will automatically add AOIs for paragraphs, sentences, words and characters, but as before, this may not be ideal depending on our research question. We can draw in specific AOIs as required, and the process is the same for all stimulus types. We select the trial where we want to add an AOI, then choose the type of AOI (rectangle, ellipse, polygon) and use the mouse to draw it onto the display. For Composite stimulus elements, AOIs are automatically included for each of the text and image components we included when creating the item.

For all stimuli, once data has been recorded a video file of the whole trial will be available in BeGaze. Time-locked or moving AOIs can be created by selecting the frame at the appropriate time point and drawing the AOI that we require. We can edit the properties of each AOI to define the start and end point of the time window when it will be applied during the trial. If we add an AOI at a specific point then add a second AOI to a later frame in a different position on the display, BeGaze can automatically calculate the movement between the two and add the appropriate AOIs to intervening frames. See Part VI, Section 7 of the BeGaze user manual (SMI, 2016a) for more information on creating AOIs for a range of stimuli.

3.2 What Are the Different Eye-Tracking Measures?

Once we have prepared our stimuli, built our study and identified the ROIs we want to analyse, we need to think about what eye-movement measures will allow us to address our research question. Eye-tracking allows us to collect and analyse data for a wide range of different measures of eye-movements and relate these to language processing and perception more generally. While we do not need to decide on the appropriate measures in advance (at least, not in terms of building them into our experiment), we should consider which ones will allow us to best address our research question when we create our stimuli.

[6] While PDFs can be static and just display a single static page, Tobii Pro Studio does treat them as dynamic (since they can be multi-page and allow a user to scroll through them), and therefore AOIs cannot be added until data has been collected.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

We discuss here the most commonly used measures in reading and visual processing studies, but other, more specialised measures may be appropriate in different contexts. For example, eye-tracking can also furnish a measure of pupil dilation, which can provide an indication of certain kinds of emotional response when reading (Laeng, Siriosi and Gredebäck, 2012). Emotions such as fear, arousal, anticipation, risk, novelty, surprise and conflict can all cause pupil dilation to increase, hence this may be of interest in studies investigating such responses.

3.2.1 Reading Measures

The literature on eye-movements during reading is well established and sizeable. We deal with reading in more detail in Chapter 4; see also Rayner (2009) for a detailed overview. We concentrate here on an overview of the measures most likely to be of use in applied linguistics studies. As stated in Chapter 1, our underlying assumption when we come to analyse eye-movements during reading is the 'eye–mind hypothesis' (Just and Carpenter, 1980). This means that what is being looked at is what is being processed, and the duration of any fixation or group of fixations is a reflection of the effort required to process what is being looked at (Staub and Rayner, 2007). Broadly, this means that in reading we assume that words that are fixated for longer require more cognitive processing, while words with shorter fixations are easier to process. Fixation times are usually measured in milliseconds (ms), although Tobii Pro Studio reports measures in seconds, and it is essential to remember that the comparisons we make are relative. In other words, longer/shorter fixations or greater/less processing effort must be in comparison to something. In our text-based study, we would be interested in the difference in the relative time spent reading the critical words 'infect' and 'affect' in our two conditions. As we suggested in Section 3.1.4, we might also be interested in the pattern of eye-movements, so do readers return to the earlier word 'violence' more in the metaphorical condition? Figure 3.10 shows both fixation and regression measures of potential interest in our example study.

As introduced in Chapter 1, eye-movement data is divided into fixations and saccades, with a complete sequence of the two being referred to as a scan path. Fixations are any

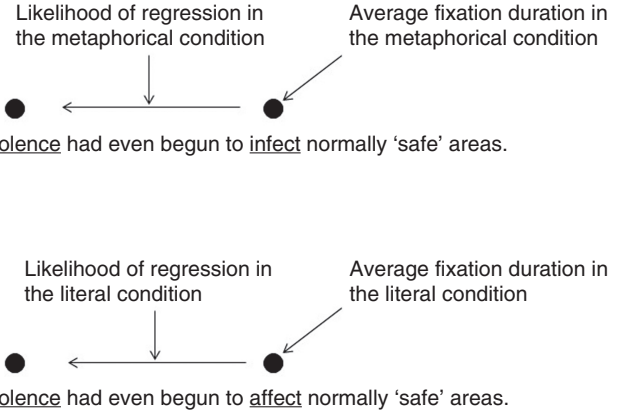


Figure 3.10 Eye-tracking measures of interest in our reading example might be whether the average fixation time on our critical word ('infect' vs 'affect') varies between conditions, and whether there is a difference in the pattern of regressions to earlier parts of the sentence.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

point where the eye is stationary on a target, and saccades are movements from one location to another. Since fixations (and specifically the duration of fixations) are more sensitive to linguistic factors than saccades (Staub and Rayner, 2007), these tend to be the measures we are most interested in for text-based studies. Saccades are important for us if we want to look at regression patterns, since longer backward saccades tend to be used to return to a prior part of a sentence in order to resolve comprehension difficulty or ambiguity. Fixation measures are classed as either 'early' or 'late' and are understood to reflect different stages of reading processing. Early measures are seen primarily as a reflection of highly automatic word recognition and lexical access processes, while later measures tend to reflect more conscious, controlled, strategic processes (Altarriba et al., 1996; Inhoff, 1984; Paterson, Liversedge and Underwood, 1999; Staub and Rayner, 2007). In broad terms, this maps onto 'first pass' measures (location and duration of fixations during the first reading of a piece of text) and 'total' measures (location and duration of all fixations, including any re-reading that is required).

Early Measures

Skipping rate is used to determine the proportion of words that receive no fixation during first pass reading and is the result of both visual factors and linguistic information. Skipped words are assumed to have been processed in the parafovea during the fixation on the previous word (Rayner, 2009, and see Section 4.1 for a discussion). Length, frequency, lexical status (function vs content words) and predictability are all determiners of whether a word is likely to be skipped. Skipping rate is reported as a probability (or percentage) and calculated as:

> Total number of trials where word was skipped during first pass reading ÷ Total number of trials (× 100 to give a percentage)

First fixation duration refers to the length of the first fixation made on a word or ROI. It is most relevant where the ROI is a single word. A similar measure that is sometimes used is *single fixation duration*, which considers only those trials where one (and only one) fixation was made on a critical word or ROI. A comparable measure that considers all fixations made on a word or ROI before the gaze exits (to the left or right) is *gaze duration*, which is also sometimes called *first pass reading time*. This provides a useful way of applying an early measure to ROIs consisting of longer words or sequences of more than one word that are likely to receive multiple fixations.

Each of the above measures can be seen as an index of lexical access, or how easily the word is recognised and retrieved from the mental lexicon. These are also the earliest points at which we might expect to see an effect of the variable being manipulated (Liversedge, Paterson and Pickering, 1998). Factors known to affect the duration of early fixation measures include word frequency and familiarity, meaning ambiguity, predictability and semantic association.

Intermediate Measures

Regression path duration, also known as *go past time*, is a measure of the time spent on the word itself and any prior parts of the sentence before the reader moves past the critical word to the right. The regression pattern can also be considered in terms of *regressions out* of an ROI (how many times a regression from the critical word to the preceding text was made – also

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

known as *first pass regressions out*) and sometimes also *regressions in* (how many times a regression from a later part of the sentence was made back into the critical ROI). These can either be measured in terms of number of regressions or as a percentage, i.e. how many trials had a regression out of or into the ROI. Regressions are hard to classify as either early or late since they can be indicative of difficulty when first encountering an item, and the subsequent time taken to overcome that difficulty (Clifton, Staub and Rayner, 2007).

Late Measures

Total reading time is the sum total of all fixations made on a word or ROI during a trial. It includes both first fixation/gaze durations and any subsequent re-reading. It is taken as a measure of the initial retrieval and subsequent integration of a word, and is likely to be affected by contextual and discourse-level factors as well as by the lexical factors already discussed. *Re-reading time* and *second pass reading time* are sometimes also reported. These measures are slightly different and have variable definitions in the literature. We define re-reading time as the regression path duration minus the first pass reading time, whereas second pass reading is the sum of all subsequent fixations on the ROI after it has been exited for the first time. (See Figure 3.11 for an example of each of these.) As well as the duration of any fixations, the total number of fixations (*fixation count*) can be measured and provides an alternative way of considering the attention paid to an ROI during the whole trial.

Later measures may not reflect purely lexical factors, and may be more influenced by contextual, syntactic or discourse-level properties of what is being read. For example, syntactic ambiguity can lead to longer fixations on a critical word or region, as well as more regressions to the prior context as the reader is forced to re-evaluate the initial analysis (Frazier and Rayner, 1982). Staub and Rayner (2007) conclude that lexical variables are the primary influence on early fixation times, while higher-level (contextual, sentence or discourse) variables are likely to show an influence later on, via re-reading and regressions and via increased overall fixation times. A crucial point to remember is that the measures are not independent: first fixation duration is a part of first pass reading time, which is a part of total reading time. Similarly, total reading time and total fixation count will generally be highly correlated. Analysing a range of measures to investigate the overall pattern should therefore be the aim, and if an effect only emerges in one of our measures, this should be interpreted with caution.

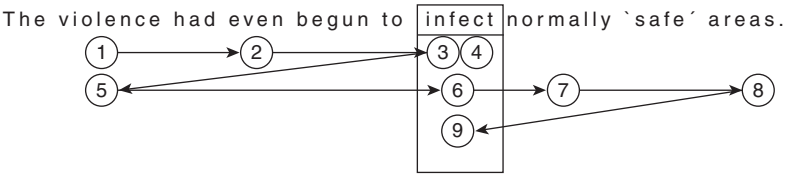
The principal reading measures are demonstrated in Figure 3.11. A range of other measures are available, and depending on our research question it might be worthwhile analysing them. For more on reading measures see Pickering et al. (2004) and Conklin and Pellicer-Sánchez (2016).

It is important to note that some of the terminology used both in the literature and in the three eye-tracking systems is not always consistent. Particularly for the different pieces of eye-tracking software, it is essential that we look at how the measures are defined in order to decide which are appropriate for our study.

3.2.2 Measures in Visual Scene Perception and Visual Search

As Holmqvist et al. (2011) point out, the sequential nature of reading does not apply to other stimuli, hence many of the measures applicable to reading behaviour are not as meaningful in the context of, for example, image- or video-based studies. Tanenhaus (2007a) makes an important distinction that in reading, longer durations are generally

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press



Processing stage and measures	Definition and depiction on figure
Early measures	
Likelihood of skipping	Number of first pass fixation durations of 0 ÷ total number of trials.
	There is a fixation during first pass [3] so in this trial the critical word was not skipped. This contributes to the total number of trials (the denominator), but not the number of skipped trials (the numerator).
First fixation duration	The duration of the first, and only the first, fixation on an ROI: [3].
Single fixation duration	Since this trial has two fixations [3, 4] on the critical word, it would not be included in single fixation duration analysis.
First pass reading time/ gaze duration	The sum of all fixations on an ROI before exiting to the right or left: [3 + 4].
Intermediate measures	
Regression path duration/ go-past time	Sum of all fixations on the ROI and any regressions to earlier parts of the sentence before moving past the right-hand boundary of the ROI: [3 + 4 + 5 + 6].
Regression rate	Number of trials with a regression ÷ total number of trials.
(or regressions out)	This example has a regression [5], which contributes to both the numerator and denominator.
Late measures	
Total reading time	Sum duration of all fixations on the ROI during a trial: [3 + 4 + 6 + 9].
Re-reading	Regression path duration for the ROI minus first pass reading time: [3 + 4 + 5 + 6] - [3 + 4] = [5 + 6].
Second pass reading time	Sum of fixations on an ROI after it has been exited for the first time: [6 + 9].
Fixation count	Total number of fixations on ROI: [3, 4, 6, 9] = 4 fixations.

Figure 3.11 Example sentence with the ROI around 'infect' indicated by a box. An illustrative eye-movement pattern is depicted below the sentence, with fixations indicated by circles and the number indicating their order. The eye-movement pattern depicted in the sentence is related to eye-tracking measures in the table below. First, the processing stage is classified as 'early' or 'late'. Second, the eye-tracking measure is given with a description of how it is calculated, as well as relating it to the fixation numbers in the example.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

equated with greater processing difficulty, while in visual-world and video studies, the measures of interest relate more to the when and where of fixations as the stimulus unfolds. Rayner (2009) provides a discussion of the differences between reading and other types of eye-tracking, such as perception of a visual scene or visual search, and concludes that there are 'obviously many differences between reading, scene perception and visual search . . . [but also] some important generalisations that can be made' (p. 1485). The difficulty of the stimuli and the nature of the task itself will determine both when and where eyes move, with more difficult stimuli leading to longer fixations and shorter saccades.

In simple image-based studies, it is a logical assumption that more looks to a specific area indicate that there is something salient or appealing about it, or that something has caused the participant to consider one part of a scene more than others. Visual-world tasks involve aspects of visual input and aural language processing, and there is a systematic relationship between what a participant hears and where (and when) the eyes move (Rayner, 2009). The same is true for video stimuli, where the participant will see what is on screen while listening to the audio track.

Visual tasks use measures such as *proportion of fixations* on various ROIs on the screen, as well as more general measures such as *total fixation times* and *fixation counts*, often time-locked to the onset of a critical stimulus. This means that we are specifically interested in how many fixations were made on a specific ROI, and how long each fixation was, during a specific time window. Measures are generally calculated relative to other parts of the display. For example, in our visual-world study with images at the four cardinal points, it is not particularly meaningful to know that there were ten fixations on the target image if we don't know how many fixations were made overall (there may have been twenty fixations on each of the other images). More useful is to compare the number or proportion of fixations for each of the four images on the screen, generally during a particular time window. The underlying assumption is that the image that is being processed when the participant hears the critical part of audio track will receive significantly more and significantly longer fixations than any of the others. A more detailed discussion of measuring eye-movements to visual scenes is provided in Section 5.1.

For dynamic stimuli the same principles apply. We would be interested in identifying how many fixations were made to specific areas of the display at specific time points in the video, or over the course of the whole video. For example, Bisson et al. (2014) measured total fixation duration, number of fixations and average fixation duration on the subtitle and image regions of the video in their study. For other studies, it is the time of the first fixation to a critical part of the stimulus that is important. Altmann and Kamide (1999) used a variant of this (*onset of first saccade*) to show that participants looked at a compatible item ('cake') earlier when they heard a biasing verb ('eat') than when they heard a non-biasing verb ('move'). Measures such as *time to first fixation* and *average fixation duration* are also of great use when we want to explore how participants' eye-movements unfold over time.

3.2.3 Selecting the Right Measures

An important part of any study is determining which measures are the most appropriate. In studies concerned with reading behaviour, a range of early and late measures are generally used, and multiple ROIs may be informative, as we have seen. The early measures described in Section 3.2.1 are most appropriate for studies concerned with lexical effects, particularly at the level of a single word. The later measures are also

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

important for analysing lexical effects, in particular how a specific word is integrated into the surrounding text, and are also useful for a study that is interested in the effects of context, or in a study that requires participants to read a longer text for comprehension. For example, if we are interested in how well language learners process newly learned or previously unknown words that we introduce in a reading task, we can use both an early measure such as first fixation duration to evaluate how easily the participant was able to recognise the word when he/she first encountered it, and a measure such as total reading time to infer how well each word was understood. Regressions to the context preceding the word might also be informative, and participants who had more trouble understanding a given word may return to the prior context more in an attempt to work out the likely meaning. If we also used a complementary measure such as a comprehension or vocabulary test, we could compare online behaviour (eye-tracking measures) with offline behaviour (test scores) as a way of further enriching our analysis.

The length of our ROIs will also determine the measures of interest. For regions longer than a single word, skipping rate or first fixation duration are of limited use. For multiword units, such as idioms, the choice of both ROIs and measures might not be straightforward. For example, analysing each individual word of an idiom might mask any 'whole phrase' effects, but only considering the phrase as a whole might also not tell us the whole story. Defining an ROI that encompasses the whole sequence, as well as ones for the component words, and using appropriate measures for each may allow a more complete picture of processing to emerge (for a discussion see Carrol and Conklin, 2014).

For image or video studies (or other dynamic stimuli), the number and duration of fixations are important, but there are different ways for us to analyse looking behaviour. We can compare raw results such as average fixation durations, overall fixation time or fixation counts, or use an alternative, like mean number of saccades, as our dependent variable (which is what Chambers and Cooke (2009) used). Other studies calculate the relative number of fixations or the relative amount of time spent fixating each ROI. For example, Holsinger (2013) compared the proportion of fixations for each of the ROIs in his study and used this as his dependent variable for analysis. It is important to understand that different variables will require different approaches to analysis, so this is something we should consider carefully in advance (see Section 7.2 for more on this).

3.3 Selecting Participants

In Chapter 1, we noted that eye-movements tend to be fairly resistant to individual differences in monolingual adults. However, the same is not true of other populations, in particular for many of the groups that are likely to be of interest in applied linguistics: children, language learners, bilinguals, and people with language impairments. General proficiency, language background, reading skill and education level will vary among participants. We therefore need to consider how to control and/or measure relevant differences when recruiting people to take part in our study (unless any of these are participant-level factors that we might want to manipulate).

If our participants are children, we need to be careful to match our sample in terms of factors like chronological age, years of schooling, reading age and language background. Children are in general less likely to sit still and concentrate for long periods, so the duration of the study is an important concern. If possible, arranging to take our equipment

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

into a school would be preferable to asking children and their parents to come to our laboratory, and would be likely to lead to more successful data collection. It is important to keep in mind that if we are dealing with children, the ethics procedures are likely to be stricter than if we are collecting data from adults.

Second language speakers will vary widely in terms of their language proficiency. We will see differences according to whether they are EFL (English as a first language) or ESL (English as a second language) learners, balanced or late bilinguals, and according to their individual experience of learning the L2 (second language). Jiang (2012) suggests that the three most influential factors in choosing participants for second language studies are general L2 proficiency, what their L1 (first language) is, and what the age of onset of L2 learning was. As well as these, we should consider specifically L2 reading and listening ability (depending on the topic of the study), and perhaps even cognitive differences like executive control,[7] which is known to be important for some of the processes of reading in a second language (Pivneva, Mercier and Titone, 2014; Whitford and Titone, 2012). Often it will be difficult, or even impossible, to find a group of participants who are matched on all of these variables. We should therefore aim to test as homogenous a group as possible (e.g. all from the same class, all with the same L1 background, all with comparable proficiency scores), then account for other individual differences in our analysis. In Section 7.2 we discuss how we might use techniques such as linear mixed effects modelling in order to do this.

Investigation of language-impaired populations is another area of applied linguistics where eye-tracking could provide a valuable tool. Such studies can be challenging since we are often targeting a fairly specific group of participants (e.g. patients with one particular type of aphasia), and sample sizes tend to be small as a result. The nature of impairments also means that our participants are likely to be quite varied, both in terms of the specific characteristics of the impairment itself, and in terms of us finding participants who all come from the same background. We should nevertheless try to match our participants on as many factors as we can. For example, if we want to use eye-tracking to investigate reading patterns in participants with dyslexia, we should aim to compare a dyslexic and control group who are matched on features like non-verbal intelligence, education level, etc.

When recruiting participants, we might also need to consider whether there are any physical or cognitive challenges that might limit their ability to take part, or which might limit their ability to provide informed consent. For some areas of study – such as dyslexia or hearing impairment – this is unlikely to be an issue. For others – people suffering from language difficulties following a stroke, for example, or older participants with degenerative conditions – we should carefully consider both the feasibility and the appropriateness of them taking part. For example, it might not be fitting to ask someone with limited mobility to come to our laboratory and sit in front of a screen for an extended period of time. It might also be the case that a person who has suffered a stroke has some degree of visual neglect that would prevent him/her from being able to view our stimuli as we intended. As with any other participant populations we are interested in, once we have recruited people who are willing and able to take part, we should do our best to account for the individual differences that are likely to exist.

[7] Executive control covers a range of functions that are necessary for the cognitive control of language and behaviour. These include things like attentional control, working memory, inhibitory control and cognitive flexibility. See Linck et al. (2014) for a meta-analysis of the role of working memory in second language processing.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

3.3.1 How Many Participants Are Needed?

This question has no simple answer, since the nature of our study will determine the type of data being collected and the amount of it that we will need. One of the main aims of most language studies is to make claims that are in theory generalisable to a larger population, for instance if we test a set of children at a certain stage of development, we would like to generalise our findings to all children at this stage of development (who are similar on other relevant factors, like years of schooling). It is therefore important to ensure that in any study we have enough participants to eliminate person-specific 'noise' or idiosyncratic behaviour (Dörnyei, 2007). Further, when deciding how many participants might be needed questions like the following are important ones to consider.

Is the study quantitative or qualitative? A qualitative study might require fewer participants but more time per person. If the aim is to investigate individual behaviour, then a small but intensive sample might be appropriate. This might be the case in studies using a think-aloud protocol combined with eye-tracking. If the aim is to conduct quantitative analysis using inferential statistics that will allow us to make claims about a participant population as a whole, then larger groups will be required to ensure that our sample is typical of that population.

How many factors will we be manipulating? We discussed this in relation to preparing our stimuli, but it is also relevant if different groups of participants are being studied. For example, we might want to compare how 'older' versus 'younger' children understand a text, or we might compare the patterns of eye-movements for adult patients with left- and right-hemisphere damage to see how this affects reading and text integration. We should bear in mind that each participant factor we introduce will effectively halve our sample size, i.e. if we have a sample of twenty children, dividing them into 'older' and 'younger' readers means we will only have ten participants per group. Remember too that in a between-subjects design with more than one presentation list, we are effectively dividing our sample size by the number of lists we have (i.e. twenty participants in a two-list study will mean only ten per list). In a two-list study with two different groups of participants, we would therefore need forty people if we want to ensure that we have ten participants per condition (ten for group 1, list A; ten for group 1, list B; ten for group 2, list A; and ten for group 2, list B).

There is no consensus on the required number of participants in an eye-tracking study, and ultimately the answer will be 'the more, the better', since the standard error for our data will be reduced as our sample size increases. As with other aspects of our study, consulting published research on a similar topic will help us to decide what an appropriate sample size might be. We should also bear in mind that some level of data loss is inevitable in any study. This can be due to technical issues such as poor calibration or computer failure, or other factors such as participants not turning up to appointments, showing highly abnormal reading patterns, etc. Planning for more participants than we actually need will allow us to mitigate some of these issues.

3.3.2 Considerations about the Participants

Video-based eye-trackers do not actually touch the eye and are safe for anyone to use (Raney et al., 2014). There are no particular risk factors that will exclude people from an eye-tracking study, although it may be necessary to identify individuals with certain conditions – such as photosensitive epilepsy or migraines – and establish whether

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

prolonged exposure to a computer screen normally has any effect on them, especially if we intend to show stimuli that contain flashing images. A rigorous screening procedure and close supervision during the experiment itself will help guard against any risk to participants.

Exclusion criteria for a study are more likely to be on the grounds of specific reading or visual disorders. Assuming we are not specifically investigating something like dyslexia, we would normally want to exclude any individuals who show 'abnormal' reading because of language or visual impairments (such as strabismus). Participants with vision that is corrected to normal with glasses or contact lenses are not normally problematic (see Section 3.4 on setting up an eye-tracker for more details), but if participants are required to attend more than one session they should not change, e.g. wear glasses to one session but contact lenses to another. Patterned or coloured contact lenses may cause problems and participants should be discouraged from wearing them to a study, and it will also help to ask participants to avoid wearing eye make-up since this can sometimes cause additional set-up difficulties. Keeping make-up remover wipes in the lab is also a good idea.

All of these considerations should be addressed when we first contact participants, to avoid anyone turning up to take part in our study who turns out to be unsuitable. Appointments should be arranged at a time when participants will be awake and alert – a drowsy participant may be both difficult to track and particularly unmotivated. As we discussed previously, for children or individuals with some kinds of impairment, taking the equipment to them may be much more fruitful and appropriate than expecting them to come to us.

3.4 Setting Up an Eye-Tracker

The physical set-up of an eye-tracker is an essential part of ensuring a high-quality dataset. Correct set-up, calibration and monitoring can ensure that our data is of good quality throughout, and minimises the risk of data loss. A large part of this comes from practice, and gaining experience as an operator of the equipment is the best way to become skilful and confident, and to learn how to deal with any problems that may arise. It is therefore a good idea to practise on a few people before starting to collect data for the first time.

Note that what we consider next are primarily eye-tracking set-ups where a participant will be invited to a laboratory to take part in our study, since these will comprise the majority of applied linguistics studies. We have mentioned some situations (such as when collecting data from children) where taking the equipment to our participants might be preferable. All three systems can be used to gather data in non-laboratory settings (see Section 2.2). We do not consider these options in detail here, but the eye-tracker user manuals provide more information. Once we have learned how to set up a stationary eye-tracker in a laboratory, adapting to a remote set-up should pose few problems.

3.4.1 Physical Set-Up of the Eye-Tracker and Participants

The physical set-up will depend on the make and model of our eye-tracker and the particular configuration chosen, but the basic procedure will be the same in most cases. The eye-tracker should be situated in a dedicated space where the conditions can remain

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

constant for the duration of a study (i.e. for all of the participants that we will want to collect data from). Consideration should therefore be given to whether a space can be found where the eye-tracker can be left set up at all times, rather than having to be cleared away or moved on a regular basis. The location should also be quiet and ideally soundproof to minimise any outside noise that could create a distraction. The space should be well lit, but not in such a way that sunlight or artificial light causes a reflection on the computer monitor or eye-tracker, with the ability to control the temperature and ventilation if possible to ensure that participants will be comfortable.

Whichever eye-tracking system we are using, a set-up guide will be provided that will help with the basics of plugging the system in, getting started, installing software, etc. The user manuals contain important information about the recommended configuration, especially the distance between participants and the monitor, which must be carefully controlled to ensure that data can be accurately recorded. All eye-trackers will have limits on the amount of head movement that can be tolerated and the size of visual angle that the equipment can record, i.e. the distance left or right that a participant can look on the screen. As a rough guide, the distance between participant and monitor should be around 1.75 times the width of the monitor to ensure that data will be accurately recorded. In practice, this means that there should be a distance of around 40–70 cm between the participant and the monitor, depending on our system and set-up. In many cases set-up and calibration issues can be resolved by adjusting the distance between participant and monitor, or by using a smaller monitor. Specific requirements and tracking limits will vary and will be detailed in the documentation received with the eye-tracking system.

The participant should be seated in front of the display screen with no distractions, hence the operator should be seated either to the side or behind the participant. In some cases, especially if the tracker is located in a small space, a partition may be useful to ensure that the participant is not distracted by what the operator is doing. The participant should be seated at a comfortable height, so a height-adjustable chair is essential. To minimise movement during a study, a non-wheeled chair is preferable. In any case, participants should be reminded to stay as still as possible during the study, with their feet flat on the floor to discourage movement. A height-adjustable desk is also very useful since this will enable us to maintain a constant set-up for the camera, screen and any chin-rest or head support. This means that once the participant is seated comfortably the entire desk can be raised or lowered as required. The participant should be seated so that his/her eyes line up around two-thirds of the way up the screen. If a chin-rest is being used, the participant's chin should be flat on the chin-rest with his/her forehead against the forehead rest. Participants should sit forward so that the forehead makes contact with the forehead rest as this will mean that the head is positioned upright relative to the monitor. Once the participant is seated and the chair and desk adjusted, we need to ensure that the eye-tracker can successfully track the eye and record the eye gaze.

In EyeLink 1000 and 1000 Plus systems we need to position and focus the camera so that the participant's face appears clearly in the centre of the display window. Once the eye is detected a green box and crosshairs will appear over the centre, and we can toggle between a view of the whole face and the eye using the left and right cursor keys. The aim is to obtain a clear image of the pupil, which will appear in blue on the display. The camera should be focused so that the corneal reflection (indicated by a turquoise circle) is as small and as sharply focused as possible. The infra-red threshold value (the sensitivity of the camera) can be adjusted up or down to ensure that the pupil can be accurately tracked. This can be set automatically, but manual adjustments may be required, for example to reduce the

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

threshold if a participant is wearing mascara or to increase it if the pupil cannot be found by the camera. Once the pupil has been found, the Search Limits option can be activated to prevent the tracker from searching elsewhere for the pupil. Various other settings (sampling rate, eye to be tracked, monocular vs binocular recording) can also be changed as required.

In both Tobii and SMI systems the set-up procedure is much more automatic. In Tobii systems, the operator does not have any control over the set-up of the camera since it is built into the monitor. The participant is seated in front of the display monitor and the eye-tracker will automatically detect and verify the participant's position and locate the eye. White circles appear on the screen to indicate that the eyes can be successfully tracked. A Track Status box helps to indicate whether the participant is seated appropriately, and again on-screen instructions help to indicate whether the position needs to be adjusted. In SMI systems, on-screen arrows indicate whether the participant is too close/too far away, and whether he/she is correctly centred in front of the display, so the position can be adjusted as required until the software indicates that the participant is sitting correctly. With both systems, once an optimal position has been obtained the participant should be encouraged to sit as still as possible throughout the experiment.

3.4.2 Calibrating the Eye-Tracker for Participants

All eye-trackers must be calibrated to each individual participant to optimise data recording. In Tobii Pro Studio and Experiment Center a calibration sequence is automatically included in each experiment, and in Experiment Builder the EL SETUP node should be included before any Trial block to incorporate the set-up and calibration sequence. Before calibrating, any major problems can be detected by asking the participant to look at the corners of the screen – if the tracking is lost at the extremities then some minor adjustments will be necessary.

The calibration itself consists of the tracker displaying a series of points on screen, which a participant must fixate in turn. The number of points used can be adjusted as required, and for studies requiring greater accuracy (e.g. full-screen reading), more points should be used. The default number of points varies in the three systems, but for most studies a nine-point calibration is standard. The calibration procedure is automatic in all three systems (i.e. the system proceeds once a stable fixation for each point has been detected) but a manual mode where the operator must accept each fixation can be used if desired. This is useful if, for example, a participant begins to anticipate the movement of the fixation points, and therefore moves to another point prior to a stable fixation being registered. The level of calibration accuracy will be reported on the operator's screen following completion, and the calibration process can be repeated as required until a high level of accuracy is obtained. All three eye-trackers include the option to replace the calibration point with an image or animation file. This can be very useful if we are working with children as it makes it much easier to attract and keep their attention to, for example, a clown or elephant than a black dot as a calibration point.

EyeLink and SMI systems also allow the operator to validate the calibration as a further check on the accuracy level of the eye-tracker. In EyeLink this is included automatically in the set-up procedure; in SMI, if a validation is needed then it must be added into the experimental structure. Since validations help confirm accuracy, it is recommended to include them whenever possible. The validation procedure repeats either the whole sequence or a subset of the fixation points and calculates the deviation of the participant's gaze from the values recorded during calibration. It will report the deviation in terms of

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

degree of visual angle. If the error for any point exceeds 0.5˚, we should recalibrate and revalidate the participant (Raney et al., 2014). In Tobii Pro Studio there is no validation option, but the accuracy of the calibration can be visually checked and verified prior to recording, and repeated if necessary. Tobii also run regular webinars on the topic of verifying data quality, and offer scripts to help with pre-study calibration that can be useful for later analysis. Once we have successfully calibrated and validated the eye-tracker, our study can begin.

3.5 Running Experiments

The process of running an experiment involves the collection of the data for each participant. This will generally include asking each person to sign a consent form in advance, setting up and calibrating the tracker, explaining any instructions, running the experiment, and collecting any demographic or language background/proficiency data that we require. Having a printed checklist prepared detailing all of the necessary steps will help to ensure that the procedure is the same for all participants and that nothing is forgotten. This is especially important if we will not be collecting all of the data ourselves.

3.5.1 Monitoring Trials

Before a study the operator should be prepared to answer any questions that participants may have about what they will be asked to do. During the study, the operator should monitor the tracker and perform any adjustments or recalibrations that may be required to ensure the integrity of the data. It is also a good idea to take notes, to keep a record of any trials that were problematic (e.g. if a participant sneezes or rubs his/her eyes during a particular trial), especially if the person monitoring the study is not the same person who will be analysing the data. This will ensure that any problematic trials can be checked and, if necessary, removed prior to data analysis. The participant should be asked to turn off any electronic devices such as mobile phones, and instructed that movement and talking should be kept to a minimum except during any breaks that may be built in to the study. Following the experiment, the operator should be prepared to debrief the participant and answer any questions.

3.5.2 Correcting for Eye-Drift

It is often a good idea to build in regular checks or recalibrations to ensure accuracy throughout our experiment. Experiment Builder also allows us to build in a trial-by-trial drift checking procedure (see the DRIFT CORRECT node that we included in our experimental structure in Figure 3.5).[8] This will display a fixation point before each trial, allowing the operator to check that the accuracy of the calibration has not changed for any reason. Once the participant has fixated on the fixation point, the system will report the visual error in the same way as during the validation procedure. If this exceeds the maximum error it will not allow a trial to begin, and will require the participant to re-fixate the point accurately before proceeding. It is up to the operator to monitor the accuracy of the drift correction point to decide if and when a recalibration may be

[8] Drift check merely checks and reports the level of accuracy, and is the default setting in EyeLink 1000/1000 Plus and Portable Duo systems. A drift correction – where the system continuously adjusts the calibration – is disabled by default on EyeLink 1000/1000 Plus and Portable Duo systems but is active by default for the EyeLink II.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

required. Recalibrations can be performed at any time, provided we have inserted the EL SETUP node at the appropriate point in the experimental structure (see the section on building studies in Experiment Builder in Section 3.1.3). Participants naturally tend to relax their position over the course of a study so small adjustments may be required over time, and a recalibration and validation should be performed after any adjustments to the set-up or after any breaks.

Tobii Pro Studio and Experiment Center do not allow for either drift correct or unplanned recalibrations, but we can build additional calibrations into the structure if required. If our study consists of multiple blocks, it is a good idea to allow participants to take a short break after each one, then for the tracker to be recalibrated prior to starting again. In Tobii Pro Studio, an experiment requiring several blocks can be constructed as separate Tests within the same Project. Each will therefore be run independently and include a calibration at the beginning by default, and the operator will need to initiate each Test separately during an experimental session. Additional calibrations can be built into the experimental sequence in Experiment Center at any point by adding Calibration and Validation elements to the stimulus list.

3.5.3 Saving Data

Following completion of the study the eye-tracking software should save the data automatically. In EyeLink the results will be saved within our deployed project folder as .edf files. In Tobii, all data is saved directly within Tobii Pro Studio and subsequent analysis will take place there. In SMI the data will be saved as .idf files inside the Results directory of the program installation directory. Ensure that all files have been saved or been successfully copied from the host to the display computer prior to shutting down either machine. If there are offline data collection aspects of our study (e.g. a background questionnaire), we should make sure that these are numbered to match the filename of our saved eye-tracking data in order to make it easy to associate the two.

In all three systems an experiment can be aborted early if required. In EyeLink experiments this is only possible during a trial, when the Abort Experiment button will be available to the operator. Tobii and SMI experiments can be stopped early using the F12 and escape keys, respectively (alternatives can be set up as required). In all cases, aborting a study will mean that any data recorded up until this point will still be recorded to the relevant participant file, but may be of limited use depending on how many items have been seen.

3.6 Conclusions and Final Checklist

We have seen in this chapter that designing, building and running an eye-tracking study involves many of the same considerations as any language experiment. As such, careful planning can help to make the process easier, and becoming confident with the system you will be using will help to make your experiment a success. We discuss ways in which eye-tracking studies can be used to address a range of research questions in Chapters 4, 5 and 6, and consider what to do with the data you have collected in Chapter 7.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

	Final checklist
• What kind of study will you be running?	⇨ Decide whether you will be using visual stimuli (text or images) or a combination of visual and auditory stimuli. ⇨ For visual stimuli, decide whether you are primarily interested in fine-grained reading behaviour (text) or broader looking patterns (visual search). ⇨ For combined visual and auditory stimuli, decide whether you will use images and spoken language (visual-world paradigm) or more authentic materials (videos). ⇨ Your research question will normally inform the methods you choose, but having a clear idea of the stimuli you will need to create will help you to plan out the rest of your study.
• What properties of the stimuli should you consider for reading (text-based) studies?	⇨ Critical words should be matched between conditions and if necessary counterbalanced between presentation lists. Features such as length, frequency and predictability will all affect reading patterns. ⇨ Critical words should not appear at the start or end of lines or sentences, wherever possible. For longer passages, critical words or sentences should not appear at the start or end of a paragraph. ⇨ The appearance of the text is important. Font size, spacing and position on the screen should all be considered. A font like Courier New (or a different monospaced font) at a minimum of 14 pt and double line spacing is recommended in most cases.
• What properties of the stimuli should you consider for image studies?	⇨ Images should be matched for size and visual salience. In visual-world studies you should consider the properties of the 'target' and 'competitor' items carefully. ⇨ Images should be positioned appropriately on the screen. The position of images should be counterbalanced so that the target image does not always appear in the same place. ⇨ Images can be created in advance or, if using authentic stimuli, you can choose specific aspects of the image.
• What properties should you consider for combined visual/audio, video and dynamic stimuli?	⇨ If recording your own audio, make sure you don't introduce any confounding variables such as prosodic cues. Audio needs to be time-locked to the stimuli to allow for accurate analysis. ⇨ Video and other types of dynamic stimuli such as webpages are likely to be authentic, so you need to choose carefully which ones to use.
• How many stimuli and participants will you need?	⇨ Methodological decisions (within-subject or between-subject design) and the nature of the study will determine this. Make sure you have enough 'power' in your study to allow you to draw clear conclusions from the analysis.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press

(cont.)

Final checklist	
• How do you build your study?	⇨ Work through the user manual and sample files provided with your eye-tracking software. Prepare your stimuli in advance then add them to the experiment. Test the experiment thoroughly before you use it to collect 'live' data.
• How do you know what to analyse?	⇨ Identify the ROIs that you will want to analyse in advance. ROIs can be added to stimuli at the build stage or added in afterward. ⇨ For reading studies, ROIs will generally be words, phrases or even whole sentences or paragraphs. ⇨ For any image- or video-based study ROIs are likely to be specific areas of the display. ⇨ ROIs can be static or dynamic. Dynamic ROIs can be set to record data only at certain points, or track an aspect of the stimulus as it moves on the screen.
• How will you know what measurements to choose?	⇨ Reading studies have a detailed literature to help you decide which measures to use. A combination of early and late measures are often chosen to allow for detailed consideration of the data. ⇨ For other types of stimuli the measures are more likely to relate to which parts of the display are fixated, or which images receive more attention over the course of a trial.
• What will you need to consider when setting up and running an experiment?	⇨ The tracker should be carefully set up to maximise accuracy. A quiet, stable location should be chosen where possible. ⇨ Calibrating and validating the tracker well will help to make sure that the data you obtain is accurate. Practice is the best way to become a competent operator. ⇨ The operator should monitor the experiment and perform any adjustments or recalibrations that he/she thinks are required.

https://doi.org/10.1017/9781108233279.004 Published online by Cambridge University Press