14 What are Eye-Movement Measures and How can they be Harnessed?

Previous chapters have described four large groups of measures which can be derived from the data produced by eye-trackers. We have also covered in detail the many ways these measures can be operationalized. Eye tracking as a research tool is particular in providing its users with well over a hundred different measures. Manual reaction time, EEG and fMRI measures come only in a handful. Why is eye tracking so different? Does this abundance of possibilities make eye-trackers particularly versatile and useful, or does it reflect some deep internal difficulty in working with eye-movement data that other biometric measurement instruments are not hampered by?

In this chapter, we investigate the concept of measures, and relate this to the methods for calculating events and representations. This allows us to propose a general model for existing measures, as well as a schema for building new ones. The chapter is divided into five short sections.

- Section 14.1 (p. 454) attempts an explanation to why there are so many measures, and why it is that researchers so often have trouble finding appropriate measures for their studies.
- The important difference between a measure concept and how it is operationalized is scrutinized in Section 14.2 (p. 456). We question the ambiguity surrounding whether the concept is referred to as a measure, or whether the operational definition itself is referred to as a measure.
- In Section 14.3 (p. 458), we more precisely relate measures to the preceding calculation of events and representations, and propose a general model for measures, based on data transformation, events, representations, and operational definitions.
- In Section 14.4 (p. 463), we argue for the division of measures into categories as outlined in the measure chapters in Part III of this book. We compare this taxonomy to other alternatives that we dismissed.
- In Section 14.5 (p. 465), we discuss different ways of enriching research with new eye-movement measures, but draw caution to the fact that developing new measures without considering those that are available already may not be a useful exercise.

14.1 Eye-movement measures: plentiful but poorly accessible

In this book, we list around 120 measures for eye-movement data. Why are there so many, and do they have to be that many? The first answer can be found in the data themselves: eye-tracking data are very versatile; rich in information in both the spatial and the temporal domains. They can be recorded in a large variety of application fields, with many types of stimuli and in combination with many other technical systems. The methods in Part II that form events and representations allow complex layers of analysis to be built.

The second reason that there are so many measures is that it is relatively easy to build new ones. Eye-movement data are very tangible, computationally easy to work with, and rich enough to generate new ideas. If a researcher is trained in mathematics or computer science, measure development and implementation is not too difficult, and may appear an intellectual challenge of the right magnitude.

While some subject areas are prone to experimentation with new measures, others can be fairly rigid. These people rely on the measures that others in their field have always used, without considering the vast range of options available to them. This may be because of lack of interdisciplinary collaboration, or a lack of proficiency in calculating new measures themselves, or simply because they work inside a paradigm or line of research where a few measures are well-examined and their effects known and trusted.

Nevertheless, current research practices allow any author of a journal paper or book chapter to define her own measure by simply writing, for instance, "In this paper, we define fixation depth to mean...". As it is much easier to define a new measure than it is to evaluate it, and because most scholarly journals put much more emphasis on the theoretical question being addressed than the measure being used to address it, journal editors and reviewers are not always expected to delve too deep into the properties of measures developed by authors. In particular, if the new measure uses mathematics that the reviewer is not acquainted with, errors in calculation and interpretation can slip under the radar. One may suspect that this is true of many new measures and the conclusions that have been drawn from their use.

Many many eye-tracking measures exist, more than we have listed in this book. Omissions are owing to two main reasons: we considered some measures not sufficiently developed to be included, and also because, despite our best efforts, it is quite likely that we have not found all of the measures that have been used.

In fact, many measures are specific to certain application fields, and only appear in their journals. Sometimes a measure is limited to a single PhD thesis, with little availability to the entire community of eye-tracking users. The fact that eye-tracking publications are so fragmented make it very difficult to find the best methods and measures. Say that you are an educational psychologist or a marketing researcher. Would you ever come across a publication in a medical imaging journal that just happens to have the measure that you need for the experiment you are in the midst of planning? You probably would not browse through all the journals where eye-tracking studies are published, because there are hundreds of them, and most are way outside your scope of theoretical interest. Search engines could do the job for you, were it not for the highly divergent terminology between the different fields of research.

Thus, somewhat paradoxically, despite there being a multitude of measures, many researchers do not find the ones that are relevant to them—sometimes it is not even apparent that a measure is relevant even when it is found. The authors of this book have many times witnessed how a researcher is unhappy with not finding the right measures for her eye-tracking experiment, and very exited when she has found a candidate. Finding the right measure is often the most difficult thing when designing an experiment, particularly for someone who is not trained in mathematics, computer science, or experimental psychology, or who is working outside established eye-tracking paradigms.

Part of the problem is that no dedicated and widely known journals exist for eye-movement methodology (as opposed to theoretical journals). No one performs benchmarking tests of eye-tracker hard- and software systems, and no standardization committees exist for hardware and data quality measurements. There is a need for systematic evaluation of the programming routines developed by the providers of eye-tracking systems: methods for filtering, detecting, and calculating fixations, saccades, and other events should be agreed upon in the industry, in collaboration with researchers. Precise operational definitions for each measure should be standardized and stated in software manuals and help functions.

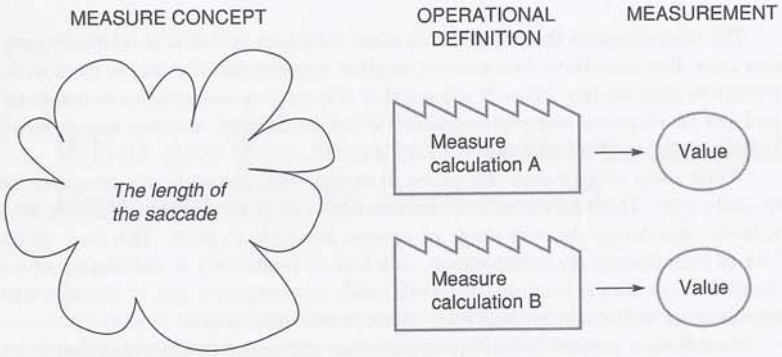


Fig. 14.1 A less precise or ambiguous measure concept, *The length of the saccade*, has two precise operational definitions that calculate the actual measure values. eye-tracking measures at their core consist of one concept and one or many operational definitions, which can produce measurement values when fed with data.

For a beginning researcher using eye tracking, the de facto authority when it comes to methods and measures has lately been the manufacturers of eye-trackers. Manufacturer terminology and the operationalizations built into their software end up in journal papers, and are often accepted without question although they should be scrutinized more.

14.2 Measure concepts and operationalizing them

Differentiating between *the measure concept*, *the way the concept is operationalized*, and the resulting *measurement values* is of great importance to understanding why we have the measures we have. A measure concept consists of the idea and name which are known in the literature and theories. Take saccadic amplitude as an example. Loosely speaking, it is the length of the saccade, which equals the distance between fixations. That is the measure concept, which is illustrated as a disconnected bubble in Figure 14.1, to highlight that the concept is mostly vague in literature as well as throughout the design of the experiment. Saccades were called "movements", and saccadic amplitude "arc of the movement" by Dodge and Cline (1901, p. 154), and it was implicitly assumed that it moved along a straight line. This implicit 'straight line' assumption about saccadic amplitudes grew into literature for many decades, and was later implemented as a calculation along the Euclidean distance between the start and end point of the saccade. Only much later were curved saccades focused upon in research, and slowly the saccadic amplitude measure became less vague but instead ambiguous. Now, in addition to the Euclidean calculation, there was an amplitude calculation given as saccade duration multiplied by average velocity (p. 311). These two methods to calculate saccadic amplitude values are the two operational definitions, and each results in its own value, which we may call the measurement. In practice, a researcher may calculate saccadic amplitude values using the operational definition implemented in her software without even reflecting that there is another one.

Even when a measure appears conceptually and operationally clear-cut, like the well-known fixation duration (p. 377) and its counterpart saccadic amplitude (p. 312), the many little known event calculation algorithms of Chapter 5 nevertheless produce fixations and saccades according to very different methods, and often this can result in differing measurement values.

A strong measure concept tends to hide variation in the way the measure is operationalized, and hence journal papers do not use labels for event detection algorithms such as 'I-DT fixation durations' or 'Eyelink fixation durations' (the operational definitions) for calculated values, but just call them 'fixation durations' (the measure concept)—although if one pays careful attention to the method section of journal papers the information about the precise calculation is sometimes provided. Few if any studies make multiple calculations of fixation durations with different event detection algorithms to validate the results they present.

When the measure concept is more vague, such as 'position similarity' (p. 359), researchers tend to produce a range of ways to operationalize it. It is not uncommon for these operational definitions to be labelled with a name and given measure status, such as the 'Mannan similarity index' or 'the Kullback-Leibler distance measure', in papers which often do not quote other operational definitions of the same measure concept. With a measure concept as weak as position similarity, we forget the overarching measure concept and are confused into believing that Kullback-Leibler distance is on par with fixation duration. In reality, the Mannan similarity index and the Kullback-Leibler distance are both operational definitions of position similarity. Just like we compare different calculations of fixation duration to one another, we should remember to compare the different calculations of position dispersion, and not let ourselves think that they are measure concepts whose importance in theory relieves us of the need to evaluate them.

The misalignment between measure concepts and their operational definitions that characterizes parts of eye-tracker-based research is a consequence of how the different measures were developed, and where they originate from. When the measure is a concept with many synonymous operationalizations (like fixation duration), it typically originates from theory in a broad sense. When precise operational definitions are marketed as measures, they are often mathematically exact solutions to specific problems in a particular line of research, and the overarching measure concept is largely anonymous. The Mannan similarity index and Kullback-Leibler distance are both examples of this.

The distinction between the theoretical measure *concept*, the *process* of performing a calculation (i.e. the operational definition) and the *value this process results in* (i.e. the measurement) is an important one because some operationalizations have been used for multiple purposes. This is true of the Mannan similarity index, a measure of position similarity which has been used to compare scanpaths, but which will tell you nothing about the temporal order of fixations, a fundamental property of a scanpath. If one is not clear on what the operational definition actually measures, i.e. what the measure concept is, much confusion can ensue and more appropriate alternatives can be overlooked.

In the choice between presenting concepts as measures or rather their precise operational definitions as measures, the authors of this book have decided to choose that which is more established. The sections of this book therefore reflect the fact that fixation duration is an established concept with several different operational definitions, while scanpath similarity, position dispersion, and position similarity are little known, and not well-established concepts in their own right. Therefore, these over-arching concepts have a set of very precisely defined 'measures' which each operationalize the concept in its own way, and which should be regarded as operational definitions.

Note that our term 'measurement' differs from 'measurement' in the sense of the totality of operations needed to establish a value, according to psychometrics and measurement theory. An eye-tracking measure has come to mean only the final calculation of, for instance saccadic amplitude, not the entire measurement completed by an eye-tracker of a specific technical build, and inclusive of all data filtering and the detection of saccades using a specific algorithm and its corresponding settings. In research with eye-tracking data, the measurement and computational history of a measure are typically forgotten, possibly because the measure

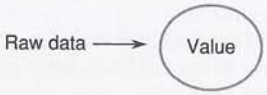


Fig. 14.2 A few measures equal a value in the raw data samples. Pupil diameter, for instance, just reads the values from the samples.

description would otherwise be too complex.

Also, the term 'operational definitions' used here differs somewhat from psychometrics and other branches of psychology. In this book operational definition is used to refer to the precise, unambiguous calculation of a value for a given measure. It thus differs from operationalization in the sense of an experimental design which utilizes an eye-tracking measure to quantify some other concept, be it fatigue, mental workload, or something else we wish to use eye tracking to investigate. For example, we could 'operationalize' fixation durations over 300 ms to indicate general interest; this is not however how the term is used to refer to eye-tracking measures in this book. 'Interest' is a possible interpretation of high fixation durations amongst others, but the operational definitions—the very calculations—of fixation duration themselves are independent of any such interpretation.

14.3 Proposed model of eye-tracking measures

By definition, every eye-tracking measure presumes a recording that produces raw data samples. Only a few measures have values that equal a value in the raw data stream, for instance the x coordinate or pupil diameter. Figure 14.2 illustrates the very simple copying of values that make up the pupil diameter measure.

Part II of this book describes methods to build events and representations by transforming the data. In Figure 14.3, we have made the method of oculomotor event detection (Chapter 5) look like a little factory to signify that it is a complex process, not unlike an industry, to take raw data of varying quality and make a refined product such as lists of fixations, saccades, and the other events. The product are events and representations, filled with more or less correct values, depending on the proficiency of the algorithms in the transformation factory.

A measure such as saccadic amplitude uses values in the events and representations, not in the raw data. Quantification of this measure requires small calculations, which we call the operational definitions of the measure. If you work with a particular operational definition you should report your measurement values according to that operational definition. There are two of them for saccadic amplitude, because this measure can and has been calculated in two different ways. These measure calculations are generally much smaller than the algorithms that transform data into events and representations. For saccadic amplitude, for instance, one operational definition consist of a multiplication of duration by average velocity, and the other of a Euclidean distance calculation between the endpoints of the saccade.

In the scanpath comparison measures, there are three levels of transformations before the measure calculation can be done. As Figure 14.4 shows, we first have oculomotor event detection, then dwell and transition detection, followed by scanpath simplification. Only after all this is done can the actual measure calculation takes place.

Based on the previous section and Figures 14.2–14.4, we can define the following:

1. An eye-tracking measure consist of a concept and one or many operational definitions. Both exist without data, like a business concept and the associated production site exist without raw material. Like any other factories, the operational definitions—implemented with algorithms—take data and refine them according to their criteria.

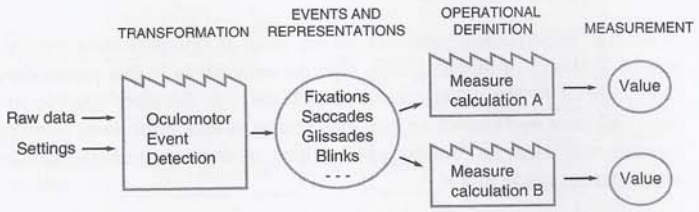


Fig. 14.3 How a measurement value for saccadic amplitude is produced from raw data and settings. The operational definition presumes a transformation of data that we know as oculomotor event detection, and which is performed using one from a variety of algorithms, using appropriate settings as secondary input (not shown). The transformation produces events and representations. The second set of algorithms (factories) are the operational definitions belonging to the measure itself, in our case the two ways of calculating saccadic amplitude, each producing its own measurement value.

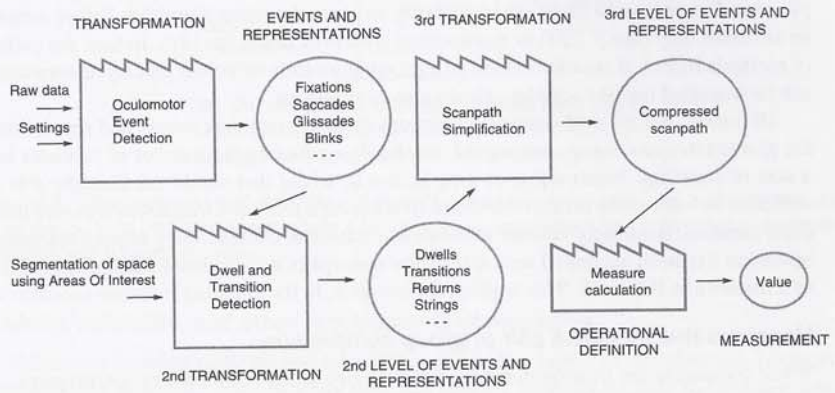


Fig. 14.4 Several of the scanpath comparison measures use repeated transformations of data before reaching the level of actual measure calculation.

2. The measure concept is expressed in words such as 'saccadic amplitude' or 'scanpath similarity' during the development of the experimental design for a study, and may have many conceptual ties to theories and previous studies. But like a business concept, the measure concept is often vague or ambiguous.
3. The value produced is called the 'measurement', and it is a product of operational definitions of the measure.
4. There are many competing algorithms (factories) on each level. In fact, choosing the right event detection algorithm for a scanpath comparison task is much like being a purchasing manager in a production site for scanpath similarity values and choosing among competing subcontractors. If you want to produce a good measurement value scanpath similarity, chose a proficient subcontractor for fixations and another for AOI strings.
5. The long calculation path shown in Figures 14.3 and 14.4 up until before the operational definition is not part of the measure. These subcontractors have to do a good job calculating fixations, saccades, and strings of AOIs, because otherwise the calculation of the value will suffer and the measurement quality will be low.

6. In general, eye-tracking measures are not taken to operationalize interest, fatigue, degree of dyslexia, or anything else. Operationalizations in this psychometric sense are taken care of by the experimental design, and are therefore specific to each experiment and must be founded on previous results in which the same interpretation of the measure was used. In contrast, eye-tracking measures are operationalized by precise mathematical calculations.

From the perspective of the researcher developing a measure, all transformation methods (subcontractors) are available simultaneously. All their products form more than 20 events and a handful of representations which can be manipulated at the operationalization stage to form a new measure. In principle therefore, any measure can be formed by combining the output from any method or combination of methods. In particular, the calculation of fixations, saccades, and other oculomotor events is *not* a prerequisite to producing other measures, which is sometimes claimed; raw sample data can in theory be taken much further. For example, you do not have to parse raw data using an event detection algorithm before generating an attention map (see p. 233) or a proportion over time graph (p. 197). In fact, the collection of methods in Part II should be seen as a growing toolbox of events and representations that can be imported into the simpler calculations of measures.

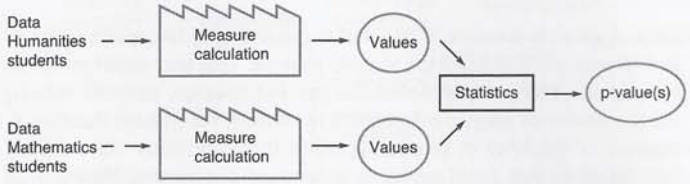
Of course, not all mathematical operations on eye-movement events and representations are granted measure status, nor should they be. Just counting the number of saccades is only a way of counting. Moreover, averaging is also so trivial that we do not consider this alone sufficient to form a new measure: average fixation duration is not a different measure than fixation duration, but simply puts the measure in a statistical context. Only when a mathematical operation has been attributed with *a measure concept* is it considered the operationalization of a measure in this book. This applies, for instance, to the counting measure saccadic rate.

Measures that include a pair or group comparison

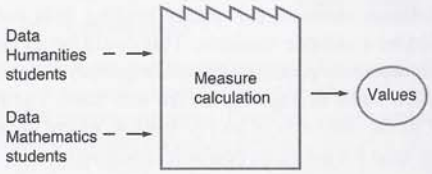
When comparing the fixation durations between two groups of data files (participants, trials, conditions...), we first calculate the fixation durations for each data file in either group (e.g. Figure 3.2). These fixation duration data are then imported into a statistics software, still in two groups of lists of fixation duration values. We tell the statistics software to make the relevant statistical comparison, which it can do, because data include all fixation duration values needed for variance calculations. Figure 14.5(a) illustrates this standard method for statistical comparisons. It gives the researcher a p-value and a df-value that can be reported in papers, and which is necessary to compute in order to estimate the likelihood that between-group effects differ from chance.

However, suppose that instead you were to compare the distribution of position values in the two data files. Here you may take the two fundamental position coordinates, (x,y), from all fixations in each group, and generate two attention maps. When comparing the two attention maps you calculate a single value which reflects the difference between them (p. 372). Neither p-value, nor df-value applies in this case. All variance information is lost. Figure 14.5(b) illustrates this.

Measures that make pair-wise comparisons are a particular case where group comparisons can be made with variance estimations. The Levenshtein string-edit measure, for instance, returns a single value representing the similarity between two AOI strings (p. 348). Again, no p-value is reported. However, when comparing AOI strings for two groups of participants using systematic multiple pair-wise comparisons between these scanpath representations, this gives a set of values that can be used to calculate significance levels (i.e. including variance information) (Feusner & Lukoff, 2008).



(a) Most measures calculate a large set of values for each condition in the experiment. These values are imported into a statistics calculation to give one or more p-values, representing the probability that we are wrong when we claim the conditions to be different.



(b) The comparison is inherent in position and scanpath similarity measures. A single value is output, reflecting the similarity between two representations.

Fig. 14.5 When two groups of data files are input to a measure, some measures both calculate and compare (b). Most measures do not include a comparison (a), which allows for a comparison in a statistics package, with p-values as output.

Validity, reliability, and other requirements of measures

Building eye-tracking methods and measures is not only a matter of computation. Measures that aspire to generality must be more than just specific methods for how to calculate some values. A number of requirements can be identified for new measures:

Usefulness First and foremost, a measure must be useful. This includes generating data which can help answer the researcher's particular question. This criterion is given. Usefulness is best proven by at least one, but preferably more experiments in a series (compare to *replication* in research methodology texts). Employing a variety of newly developed eye-tracking measures, Goldberg and Kotval (1999) highlighted which ones are appropriate for identifying differences between poor and good computer interface design. Further comparisons of measures between conditions should be encouraged. Meta-analyses, moreover, provide insights into the usefulness of measures. Meta-analyses often focus on the validity and reliability of specific interpretations of measures. For example, O'Driscoll and Callahan (2008) in schizophrenia research, Clifton *et al.* (2007) in reading research and Jacob and Karn (2003) in usability research.

Uniqueness A new measure should differentiate from earlier ones in a relevant way, in the way it deals with the data, reduces the information, and summarizes it. There is little point in having yet another measure which provides values for something we can already measure perfectly well given already existing measures. Uniqueness has been difficult to demonstrate due to the fragmentation of eye-tracker-based research. For instance, the many position dispersion measures were often proposed without any demonstration of uniqueness. As alluded to before, this fragmentation stems from a lack of clarity in what is actually being measured.

Reliability A reliable measure is one that is consistent in its measurement of the same effect in contexts it is expected to generalize to. This may entail reliability across time, participants, and/or experimental designs. For instance, saccadic velocity may be a reliable measure of fatigue independent of context, but fixation duration is *not* a reliable measure of the level of processing, since there are many other causes of long fixation durations that could appear in almost any experiment. Note that an eye-tracking measure can be reliable within only *one specific paradigm or type of experiment*. In standard reading research, for example, first fixation duration appears to be a reliable measure of lexical processing because this has been established over many experiments. By contrast, if we go to a completely different branch of research in reading with much different material, e.g. poetry reading, it is not obvious that first fixation duration would be a reliable measure. This could be so because poetry reading does not necessarily strive to process a word as quickly as possible and then proceed, as there is an added value in lingering on the text units. The reliability of a measure can be thought of as the precision of an eye-tracker, where the same measurement value is produced over time for the same position/conditions.

Validity A valid measure is one that measures what it is intended to measure. What we intend to measure, however, usually boils down to two things. First, a measure can be valid with regard to the oculomotor event it measures. For example, the fact that a fixation duration actually measures the durations of an objectively determined event known as a fixation. Secondly, a measure can be valid with regard to the cognitive process we believe is involved in generating this eye-movement. For example, that the duration of a fixation on a text unit actually reflects the processing difficulty for the reader. Eye-tracking measures may or may not be open to interpretation within specific paradigms and experiments, and the interpretation may or may not be valid. The validity of implementing a measure for a specific purpose can be assessed in a number of ways. Concurrent and predictive validity can be established by correlating the proposed measure with a criterion measure which is known to be valid. For instance, one could correlate a psychiatric diagnosis of schizophrenia with various smooth pursuit measures. Content validity refers to the degree to which a measure covers all variation of the effect (i.e. is the number of catch-up saccades in smooth pursuit higher for all varieties of schizophrenia?). We can think of validity as the accuracy of the eye-tracker: we may still have reliable measurements over time (precision), but indicating the wrong position (invalid). Getting the correct position corresponds to us having a valid measure to answer our research questions.

When referring to validity and reliability in the last two points, we do so in the sense of what the measure can be used for, irrespective of eye-tracking terminology. Of course, there is also the question of whether the measure encapsulates the concept properly, the *concept validity*. The *internal validity* of an eye-tracking measure could be taken as an index of, for example, how well operational definitions of saccadic rate or scanpath similarity actually quantify the frequency of saccades, or the degree of resemblance between eye-movement sequences, respectively. Other types of validity also exist that do not exclusively relate to measures themselves. For instance *ecological validity*, which refers to the ability to generalize the result; that is, how similar the whole experimental situation and procedure is to all other situations that the study purports to shed light upon. What do results from a car simulator study tell us about real car driving, for instance. Figure 14.6 illustrates the different types of validity against a dimmed background of Figure 14.1.

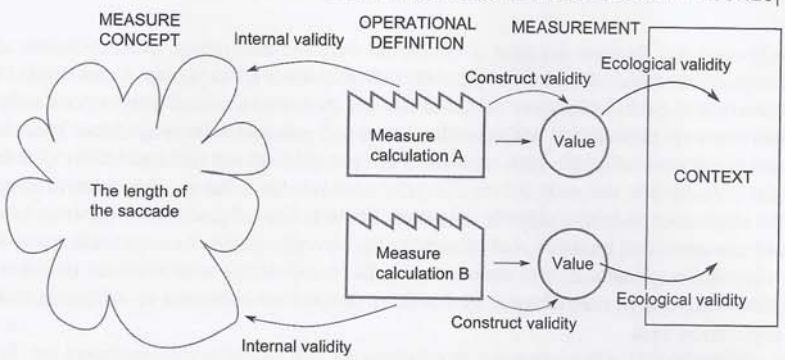


Fig. 14.6 *Internal* validity refers to whether the operational definition actually measures the concept, for instance does your measure of saccadic amplitude actually measure saccadic amplitude? Internal validity is more important for more complex measures. *Construct* validity concerns whether the operational definition actually measures the intended behaviour, such as, does your measure of saccadic rate actually measure tiredness? *Ecological* validity is whether the operational definition and measurement value obtained in the recording scenario generalizes to the real world. Is it a valid measure in other contexts outside the Lab or wherever else the test was carried out?

14.4 Classification of eye-movement measures

In journal papers and book chapters, on the internet, and in some dissertations, it is possible to find lists of eye-movement measures and accompanying interpretations. There have been very few attempts however to classify measures of eye-movements across research fields according to a broad taxonomy. The nature of measures becomes even clearer when we compare different classification systems.

In this book, we have used a principle of categorization built upon *operational definitions*. Take the measures of latency in Chapter 13 as an example: any measure that is defined as a duration between two separate events will be found in the latency chapter, whether the two events are the stimulus onset and the start of the saccade, the time it takes the eye to reach an area referred to in speech, after speech has commenced, or the duration between a puff of air on the eyeball and a resulting blink. We considered that using operational definitions from published research as a guiding principle to build measures into a classification system was the optimal approach. We hope that readers in that stage of an eye-tracking study where operational definitions are key, and measure selection is crucial, will see the benefit of structuring measures in this way. The semi-encyclopedic nature of Part III means that we assume that anyone reading about a particular measure is interested in using it. We do not explain *when* or *why* it is important to consider latency measures; we assume that the reader can decide that for themselves, comparing the information provided along with each measure to the needs of their own experimental design.

One tempting and common suggestion is to divide measures by *application field*. Comprehensive lists and meta-studies have been made of 'human factors measures' (Rötting, 2001), 'car driving measures' (ISO/TS 15007-1, 2002; ISO/TS 15007-2, 2001), 'usability measures' (Jacob & Karn, 2003), 'reading measures' (Clifton *et al.*, 2007), and 'schizophrenia research measures' (O'Driscoll & Callahan, 2008) etc. Some measures have indeed been tightly associated with a particular paradigm in a particular field of research. Proportion over time graphs are mostly found in psycholinguistic research, anti-saccade measures in neuropsychological studies, and first pass regression scanpaths are mainly limited to papers on reading. However, measures are more closely related and relevant *between* fields than researchers think. Whilst

it is clear that fixation duration is ubiquitous between disciplines, other measures are being adopted more and more by people with different areas of expertise, for example, the Levenstein string-edit technique. This measure has its roots in mental imagery research, but has subsequently appeared in both usability and scene perception investigations. There are many more examples of where such transitions are possible: do not feel restricted to use measures just because it is common practice in your research field, but use this book to shop around for alternative and more suitable measures. There is a lot of potential in implementing different measures and methods, but remember that novelty alone does not make good research; *your choice of methods and measures should be tailored to your research question and experimental design* (see Chapter 3). For these reasons we chose not to categorize methods by application field.

If we would adhere to strict psychometrics, we would divide measures by their *interpretation*. A psychometric division would give categories of 'early processing measures', 'prefrontal control measures', 'emotional reaction measures', 'interest measures', 'workload measures' etc. However desirable such a division might seem, there are serious problems with it. As interpretations of measures vary between paradigms, measures would appear in more than one category, and have no single natural place in the system. Moreover, eye-tracking data do not have straightforward interpretations outside of the context of the experiment where they were recorded. Thus we could not say conclusively for instance, that longer fixation durations *always* mean that the fixated region has been processed to a greater extent. Furthermore, interpretations of lesser known measures are far from validated; some have appeared in only one journal paper and never been used again. The recent commercialization of eye-tracker-based studies has led to dubious interpretations of eye-tracking measures to say the least. Meaning is attributed to some measures in a manner which appears purely subjective—heat maps (p. 239) and pupil diameter (p. 391) in particular, are often taken to indicate constructs such as 'informativeness' and 'mental work load' respectively, where more likely causes of the effect have not been accounted for. Categorizing measures by their interpretation may thus inadvertently support an overly simplistic view of eye-movements and what they actually reveal insights into. Also, new research revises and finds additional interpretations all the time, so such a division would soon be outdated.

The precise calculation of measures depends on the computation of events and representations in Part II of this book. One alternative means of measure classification would be to create a taxonomy based on this, i.e. *the preceding calculations of events and representations*. This would give us categories such as 'fixation measures', 'AOI-measures', and 'scanpath measures'. However, although many measures indeed take only one type of event as input, many others combine two or three events, or even make use of representations to build a measure: number of fixations in AOIs for instance, the main sequence measure, and transition matrix measures. Moreover, an event-based division would also obscure the large similarity between some measures, such as fixation duration and dwell time, which would end up in different chapters according to an event-and-representation-based classification system.

A similar idea would be to categorize measures according to *how many steps need to be taken with data* to reach the measure? Karn, Ellis, and Juliano (1999) propose a categorization of all measures into four levels, similar to that shown in Figures 14.2–14.4 on pages 458–459 above. Pupil dilation comes right out of the gaze estimation algorithm, while fixation durations require additional filtering and event calculation. Heat maps add a layer of Gaussian calculations, and the Levenshtein string-edit measure adds AOI segmentation of the stimulus, and data alignment. Every measure has a calculation path from raw data to measure value, but how informative is it to the reader to have chapters entitled 'Measures that can be directly used from raw data', 'Measures that require event detection', and 'Measures that require AOIs'? Again, measures with large similarities would end up in different categories, according to this

structure.

Lastly, one could potentially classify measures by *mathematical principles*; this would emphasize the crucial calculations behind measures, the very foundations of the numerical values we report as scientific results. More mathematical and information theoretic tools are gradually diffusing into eye-tracking research, and the space for mathematically tractable measures will slowly be filled by new research to the extent that it is needed. A categorization of chapters according to mathematical principles would require chapter titles such as, 'Measures that require trigonometry' and 'Measures that are based on correlational analysis'. Whilst this could doubtless further the detailed understanding of measures, such a book would be less accessible to many eye-tracker users, as hard-core mathematics can be intimidating to some.

These principles for classification of measures all describe important properties across all measures, but for the reader who uses the book to find methods and measures at the point in time when designing an experiment, we consider a classification system based on operationalizations to be the most efficient—this is how Part III is organized.

14.5 How to construct even more measures

Eye-tracking data have a structure that lends itself to many types of data analysis, which is one reason why we find so many measures. Suppose that we want to enrich eye-tracker-based research with even more measures—how should we go about this?

First of all, it is important to realize that *analysing eye-movement data is always a matter of reducing the amount of information in the data*, and through that reduction one achieves overviews and summaries at a new level of information. Reducing and selecting information is what the data transformation methods of Part II do for us. A recording, with its raw data samples and stimulus images is so full of unstructured information it is virtually impossible to make sense of without some form of reduction and re-representation; conversely, a transition matrix with its simple summary does not retain very much of the original information from a recording. Hence, devising a new measure is largely a matter of actively deciding what information *not* to keep, rather than just what summary to produce. This can be a fine balance.

Of course, it is the research question and the experimental design that you are working on which decide what summary of data your new measure should produce. If your hypothesis is that fixation stability is better for professional elite shooters, then you need to worry about what to do with the raw data samples inside fixations. Is the size of a boundary box a good enough summary, or does your experimental design require that you quantify each minute intra-fixational eye movement for you to prove that your hypothesis is correct? The choice of precise operationalization depends on the counter-arguments you can expect to your conclusions, and this depends on the theory and earlier studies on fixational eye-movements that you would have to consider. From this perspective, Viviani (1990) was correct when concluding that eye-movement data can only be interpreted within a specific theoretical framework.

It should however be emphasized that it is equally possible to construct measures from fictitious data alone, using the existing events and representations, without having a particular experiment or theory in mind. For instance, take the measure we called angle between dwell map vectors (p. 374). It compares two dwell maps, that is the dwell time values in the cells of two gridded AOI matrices, and tells us how similarly distributed they are. Now, transition matrices also consist of values in the cells of a matrix. Mathematically, one can use the procedure from the measure 'angle between dwell map vectors' to compare two transition matrices rather than two dwell maps. This is perfectly valid in mathematical terms. The output value

would tell you how similar two transition matrices are with respect to which types of transition are common. In other words, we can indeed interpret this new measure—which we may call 'the angle between transition matrix vectors'—without ever designing or conducting an experiment, or thinking about a particular theory.

The following list presents six examples of how new and useful measures can be invented:

Detect a new event in the raw data samples Microsaccades and glissades are examples of events that have only become detectable relatively recently, owing to the advancement of event detection algorithms and faster sampling frequency of eye-trackers with better accuracy and precision. It may seem that there is not much room for further events in the raw data stream, but this is clearly not the case. Take the multi-step saccades on page 327 as an example. Detecting and accurately measuring these could yield a variety of velocity and counting measures. Another possibility would be to use thresholding on pupil data, for instance to identify all increases in pupil diameter lasting for at least 100 ms, with a total increase of at least 0.3 mm, and no intermediary decrease larger than 0.05 mm, and count the number of such occurrences in your trials.

Define new events from AOI strings and scanpath representations In-word and between-word regressions, and first pass regression scanpaths are only a few of the full possibilities of events which could be constructed from scanpath or AOI-based representations. Hyönä *et al.* (2003) propose several innovative events for the analysis of reading data according to such principles. Mathematically there is an infinite number of scanpath forms, and more than a handful could be expected to be generally useful. There is a great need for such measures in several of the applied fields; usability and educational psychology, to mention a few.

Re-use representational forms across types of data The similarity of dwell maps to transition matrices allows us to reuse the angle between dwell map vectors measure for transition matrices. Seeing similarities, breaking hidden assumptions, and generalizing a measure can make it applicable to new types of data. As another example, take the AOI dwell string. All measures assume that these are strings of AOI names, but from a computational viewpoint, a string of fixation durations or a string of saccadic amplitudes makes just as much sense, and such strings could easily be processed using the AOI string and scanpath measures we describe on page 339.

Import mathematical constructs Mathematical constructs can be imported that are similar to representations in eye movement data. There are many previous examples: string-editing stems from methods for comparing gene sequences, which were seen as similar to sequences of dwells. Scanpaths are similar to sequences of vectors and graphs. Gridded AOIs are matrices covering space. Mathematics is large in scope however, and there are many untried candidates. Time series has obvious similarities with both raw data and all sequence and over-time data. Knot theory deals with graphs that have a great similarity to scanpaths.

Plot a single-value measure over time If a measure can be calculated to give a value per data sample, or per fixation or any other chronometric unit, then it can be plotted over time, and the growth and decline of this measure can be modelled. Both the magnitude and the time course of the effect driving the measure can be examined. In fact, a large number of the measures in earlier chapters can be made into a value-over-time measures, when restricted into a time window and used across many participants, stimuli, or trials (according to the experimental design). For instance, the value-over-time measure 'Average fixation duration over time' is produced by selecting a time window of, say 500 ms, and for all participants and/or stimuli, averaging the durations of all

fixations starting in each interval. This will give a curve showing how the average fixation duration changes over a trial. The measure 'Dwell time over dwell order' for each dwell, in order from one and up, sums the dwell times across participants, stimuli, or trials. The measure 'Number of returns over time' counts number of returns to AOIs already seen in successive windows of size 1 second, and could be expected to vary during decision processes, for instance in supermarket shopping (see Gidlöf *et al.*, in progress). These value-over-time graphs are easy to produce, and if the data files are not too long, they can even be calculated in a spreadsheet program such as OpenOffice Calc.

Replace or reverse space and time By replacing spatial dimensions with feature dimensions (p. 215), many measures can be revived in new forms. Reversing and binning time (p. 205) allows for a whole array of measure variants.

Amalgamate measure concepts In the context of some experiments one or either measure that you choose as a tool to tackle your research question, may not be sufficient in its own right. In such cases it is theoretically possible to merge measures. For instance, suppose you expect participants to 'look for longer' on a particular AOI in one condition of your experiment compared to another. Although this prediction is imprecise it poses a genuine problem in that the number of fixations and their duration might confound each other with respect to your hypothesis. If one condition has a larger number of fixations on the AOI than another, in the predicted direction, but the duration of fixations between conditions is comparable, it is not possible to conclude that people 'look for longer' only that they 'look more often', since conventional statistical tests rely on the calculation of means and variance. The mean fixation duration on the AOI will be comparable between conditions, but fewer data points will contribute to it in one condition. The latter conclusion, 'look more often', thus relates only to frequency, despite the fact that the sum total of fixation durations is longer in the condition that you predicted it would be. Dwell time is not useful since some trials contain no fixations at all on the AOI in question. Again this is in line with the hypothesis, but not captured by your two measures (number of fixations in an AOI, and mean fixation duration in an AOI). One possible solution therefore, is to combine the two measures and take an average of fixation duration for each condition, *including trials in which no fixation occurred* (Dewhurst & Crundall, 2008; Dewhurst, 2009). The resulting value will be an amalgamation of number of fixations and fixation duration, and should be higher in line with your prediction. This issue relates to missing data, and the calculation of measures with respect to the experimental design and how eye-movement data is collapsed across trials and conditions (p. 78). The principle of merging measures in this way however, could be applied to a range of different events and representations, where the field of mathematics is open to you.

On the one hand, a systematic investigation into the full possibilities of developing measures for eye-movement data would be of great theoretical interest. It must be properly grounded in the needs of actual research, however, so that measures devised can actually come to good use. So far the mathematics used within measures is somewhat dichotomous: many measures rely on simple event counting, AOI region inclusion calculations, and possibly some trigonometry; others involve advanced filtering, probability density functions, Markov models, multidimensional vector spaces, and skewness and entropy calculations. This gives the impression that one half of the measures are only for mathematically advanced researchers. This is particularly paradoxical, in view of the fact that many applied researchers are in need of more mathematically advanced measures, while many researchers in neurological and vision research fields are satisfied with event counting and region inclusion calculations.

It is necessary to draw caution to the development of new measures without considering those that are available already, however. This may not be a useful exercise. A vast choice of measures are packaged and ready to go, being well grounded in theory and empirical investigation. Developing a new measure should really only be a consequence of encountering a specific problem, and even then you should be sure that you know the implications of your new calculation for your results. For the inexperienced researcher a likely reason for wanting to calculate a new measure could be that the experimental design is not tight enough, or some aspect was overlooked. Moreover, if we calculate too many measures, the fragmentation between research fields will increase exponentially, raising more of the problems already addressed in this chapter, most importantly the difficulty in deciphering what the measure actually measures.

14.6 Summary

There are so many eye-tracking measures, because eye-tracking is so versatile, and it is easy and fun to construct new measures. Easier, perhaps, than to find an already existing measure in the highly fragmented channels for publication of eye-tracker-based studies. In fact, the number of methods and measures for eye-tracking grows much faster than the validation of their interpretations. Awaiting enough empirical evidence, the responsibility for using measures correctly therefore rests upon authors and reviewers. Standardization of existing measures with respect to terminology and precise operationalizations is of essence to support them.