Behav Res (2018) 50:1645–1656
https://doi.org/10.3758/s13428-017-0998-z

Threats to the validity of eye-movement research in psychology

Jacob L. Orquin[1] · **Kenneth Holmqvist**[2,3,4]

Published online: 7 December 2017

Abstract Eyetracking research in psychology has grown exponentially over the past decades, as equipment has become cheaper and easier to use. The surge in eyetracking research has not, however, been equaled by a growth in methodological awareness, and practices that are best avoided have become commonplace. We describe nine threats to the validity of eyetracking research and provide, whenever possible, advice on how to avoid or mitigate these challenges. These threats concern both internal and external validity and relate to the design of eyetracking studies, to data preprocessing, to data analysis, and to the interpretation of eyetracking data.

Keywords Eyetracking · Best practice · Experimental design · Data analysis · Researcher degrees of freedom · Internal validity · External validity

Eye-movement recordings began in the 19th century. During most of the 20th century, it was very difficult and expensive to record and analyze eye movements. Researchers who built or bought an eyetracker could easily spend a year setting it up, and the analysis was equally time-consuming. Hartridge and Thomson (1948) devised a method for analyzing eye movements at a rate of almost 3 h of analysis time for 1 s of recorded data, and as Monty (1975) remarked: "It is not uncommon to spend days processing data that took only minutes to collect" (pp. 331–332). Even in the 1990s, eyetrackers were found in only a few psychology, biology, and medical labs, at places such as NASA, and in some very tech-savvy commercial advertisement companies or car manufacturers. Usually there was enough time to acquire the method from knowledgeable colleagues and to run numerous pilots before the actual data were recorded and analyzed. Since the early 2000s, eye-movement research has been adopted in many new disciplines, many of them applied and full of researchers with little experience in experimental design and statistics. This diversification of eye-movement research has largely been driven by technological development: Modern video-based eyetrackers drastically simplified eyetracking, often with a "plug-and-play" approach. Some of the eyetracking hardware companies were highly successful in expanding their customer base into new areas by making eyetracking seem easy. Although the eyetracker users extended into new fields, the experimentation and analysis skills necessary to operate the equipment did not always follow suit. For example, a survey of eyetracking research on decision-making (Schulte-Mecklenbeck, Fiedler, Renkewitz, & Orquin, 2017) showed that 35% of the reviewed studies included fewer than 16 critical trials. The reviewed studies originated from various disciplines, such as psychology, marketing, economics, neuroscience, and human–computer interaction. The same survey showed that 20% had fewer than five trials, and 12% had but a single critical trial (Schulte-Mecklenbeck et al., 2017). Although a single trial might be standard in medical research, it is rarely recommendable in eyetracking studies using, for instance, naturalistic stimuli. In this article, we caution against using such a low number of trials (see the Undersampling of Naturalistic Stimuli section), since it diminishes stimuli

✉ Jacob L. Orquin
jalo@mgmt.au.dk

[1] Department of Management/MAPP, Aarhus University, Fuglesangs Alle 4, DK-8210 Aarhus V, Denmark

[2] UPSET, North-West University, Vaal Triangle campus, Vanderbijlpark, South Africa

[3] Faculty of Arts, Masaryk University, Brno, Czech Republic

[4] Department of Psychology, Regensburg University, Regensburg, Germany

representativeness and threatens the external validity of the study. The survey also reveals that many studies use total dwell time as a dependent variable and that many studies analyze multiple eye-movement metrics (see also von der Malsburg & Angele, 2017). Here we advise against the use of total dwell time (in the Total Dwell Time section) and against analyzing multiple eye-movement metrics (in the Analyzing Multiple Metrics section). We consider the former a threat to the construct validity and the latter a threat to the statistical validity of eye-movement research.

Motivated by these concerns, we outline a number of threats to the validity of eye-movement research. Shadish, Cook, and Campbell (2002) have described a general list of threats to the validity of experimental and quasi-experimental research. Following their example, we organize our list into threats to internal and threats to external validity. By *internal validity*, we refer to the extent to which warranted, and sometimes causal, inferences can be made from eyetracking studies, and with *external validity*, we refer to the ability to generalize these inferences to new populations and stimuli.

Throughout the article, we refer to various studies to illustrate different points about eyetracking research practices. It is important to note that although some studies are used as examples of practices that involve threats to validity, each study must be understood in its own context. In experimental design, we are often forced to make trade-offs between various problems and threats. When solving one problem, we often acquire a new one. If we, for instance, use simplistic stimuli to achieve internal validity, we often sacrifice external validity, and vice versa.

We do not wish to reiterate what has already been said about the proper way to conduct eyetracking research (for overviews, see Duchowski, 2007; Holmqvist et al., 2011; Russo, 2011), but hope to challenge common assumptions in eye-movement research and to increase awareness of methodological pitfalls. Although we believe that all threats are described in sufficient depth to make recommendations for eye-movement research, our examination is far from exhaustive.

Threats to internal validity

Inappropriate comparisons

Many eyetracking studies aim to compare the distribution of eye movements to different objects in an image. For instance, Dodd et al. (2012) investigated whether participants fixate more pleasing or more aversive objects, depending on their left-wing versus right-wing political orientation. Glöckner and Herbold (2011) studied whether decision-makers fixate more on the probabilities or the payoffs when choosing between risky gambles, and Baker, Schweitzer, Risko, Ware, and Sinnott-Armstrong (2013) studied whether readers of neuroscience articles pay more attention to neuroimages than to bar graphs. Although these examples may seem uncontroversial, the last example is, at least in principle, an inappropriate comparison. In the first example (Dodd et al., 2012), comparisons are made between groups of participants with respect to the same stimuli whereas the last (Baker et al., 2013) compares between stimuli (neuroimages vs. bar graphs). Contrary to the authors' expectations, readers pay less attention to the interesting neuroimages than to the supposedly dull bar graphs. Why could this be an inappropriate comparison? The possible causes for fixating either object differ. Bar graphs could very well receive more fixations than neuroimages because they are harder to understand, not because they are more interesting (Shah & Hoeffner, 2002). The risky gambles example can in principle lead to a similar challenge. Suppose, for instance, that a study predicts that participants use a decision strategy that results in more fixations to payoffs than to probabilities. In experiments with gambles, information is typically presented using the same number of characters—for example, "15%" and "$25"—but imagine that payoffs were presented as "twenty five dollars." If so, participants would need more fixations and longer time to process the payoff information because of its unfamiliar presentation and the fact that it contains 19 rather than three characters (Rayner, 2009). Such a presentation would lead to a difference in eye movements in the predicted direction and we would wrongfully conclude that the data supports our prediction. Even in the standard case in which probabilities and payoffs are presented using numbers, one could make a similar argument that the lower familiarity of probabilities could lead to longer fixation durations. The problem with inappropriate comparisons is particularly unfortunate considering the aim of much eyetracking research—namely, to compare eye movements executed to different stimuli. There are, however, a few ways of solving this problem:

- The researcher examines differences in eye movements due to stimulus features and develops or selects stimuli that differ systematically on one or more features (see, e.g., Orquin & Lagerkvist, 2015; Towal, Mormann, & Koch, 2013).
- Comparisons are made between different groups of participants to the same stimuli. Dodd and colleagues, for instance, compared whether political left- versus right-wing participants fixate more on positive or negative images thereby avoiding a direct comparison between different types of images (Dodd et al., 2012).
- The comparison is made between sets of stimuli that are large enough to assume that irrelevant feature differences randomize away (see the section on Undersampling Naturalistic Stimuli). Nummenmaa and colleagues, for instance, compared 16 pleasant to 16 unpleasant and 16 neutral images to understand attention capture by aversive

stimuli relative to positive or neutral stimuli (Nummenmaa, Hyönä, & Calvo, 2006).

Analyzing multiple metrics

Recognizing data fishing in psychology and attempts to counter it are becoming more commonplace (Wicherts et al., 2016), but what about eyetracking research? As it turns out, eyetracking research probably provides an even higher number of *researcher degrees of freedom* than other quantitative methods. Eyetracking data requires multiple preprocessing steps and each step can be adjusted to provide a different result: Changing the size of areas of interest (AOI) can, for instance, improve the fit of a model (Orquin, Ashby, & Clarke, 2016). A surprisingly common feature in eyetracking studies is comparison of multiple AOIs on multiple eye-movement metrics (von der Malsburg & Angele, 2017). For instance, in a study on food nutrition labels, Antúnez et al. (2013) compared six AOIs in one condition and four AOIs in another on five different metrics yielding 105 significance tests. In the absence of a Bonferroni correction or directed hypotheses, it makes no sense to interpret these significance tests. Another challenge with this approach is that the metrics in question tend to be highly correlated, such as total fixation duration, fixation count, and visit count.

Perhaps this highly data-driven approach to research has become popular because the data processing tools from commercial vendors invite their users to try out a broad scan of all possible comparisons. Although exploratory approaches have their merits, most eye-movement studies would benefit from directed hypotheses and predictions. Fortunately, it is easy to avoid analyzing multiple metrics by following a few simple steps: (1) Formulate a hypothesis from theory, earlier studies, pilot studies, or lay notions, and think of it in terms of eye movements. (2) Take the stimulus or trial mechanism and draw or simulate participants' expected eye movements. (3) Consider what is most important in the drawing or simulation in order to test the hypothesis: movement, position, latency or numerosity measures? (4) Finally, consult a list of measures (e.g., Holmqvist et al., 2011), and settle only on those measures necessary to test the hypothesis.

Data quality

Data quality comprises many aspects of research—for example, the end-to-end latency (Reingold, 2014), tracking loss, or sensitivity to a participant's movements (Niehorster, Cornelissen, Holmqvist, Hooge, & Hessels, 2017). Data quality can vary considerably across eyetrackers. The average accuracy (validity) ranges from around 0.4° to around 2° (Holmqvist, Zemblys, Mulvey, Cleveland, & Pelz, 2015). The difference in precision (reliability) has even a larger range, from around 0.005° root-mean squared (RMS) in the best remote eyetrackers, to 0.5° RMS in the poorest (Holmqvist et al., 2015). These data quality issues imply that fixations are never measured at their true location begging the question of how small objects can reliably be studied with eyetracking. For instance, using a Tobii eyetracker with a presumed accuracy of 0.5° and precision 0.35°, Donovan and Litchfield (2013) studied detection of cancer nodules, the smallest of which were 0.28°. Similarly, Orquin and Lagerkvist (2015) studied detection of product labels that were 1.8° using a Tobii eyetracker with an accuracy of 0.5° and precision of 0.18°. In both cases, the obvious question is whether the stimuli are large enough for the respective eyetrackers. So far, no standard to determine the smallest possible object that can be used with a given eyetracker's accuracy and precision has been proposed.

In order to propose a standard, we introduce a few concepts. We refer to the percentage of fixations to an object that fall within the boundaries of the object as the *capture rate*. Low capture rates may cause several problems such as uncertainty about the amount of fixations to a given object, and if objects are close to each other, it leads to assignment of fixations to wrong AOIs (Orquin et al., 2016). The capture rate is a function of the true location and distribution of eye fixations and the hardware-related noise distribution. If the properties of the true fixation distribution are unknown, it is safest to assume that fixations are uniformly distributed within the boundaries of the object, thereby making no assumptions about which parts of the stimulus are more likely to be fixated.

To understand the different factors that may influence the capture rate, we perform a simulation study on the effects of accuracy, precision, stimulus size, stimulus shape, offset angle, and the centrality of the fixation distribution. We examine the effects of accuracy, precision, stimulus size, and fixation distribution separately, and the effects of stimulus shape and offset angle together. Unless stated otherwise, the simulation assumes a round object with the true fixation locations uniformly distributed inside the object. All simulations follow the same procedure: First, we obtain the true fixation location by drawing 100,000 random samples from a bivariate uniform distribution. The distribution ranges from (0, 0) to (x_{ul}, y_{ul}), where x_{ul} and y_{ul} are the upper limits on the x- and y-axes. We then retain all fixations that fall within r° of the center of the distribution, thereby obtaining a circle with r being the radius. Then we draw offset angles uniformly—that is, the direction in which the fixation is being offset, between 0° and 360°—as well as offset distances from a normal distribution with mean equal to the accuracy of the eyetracker and standard deviation equal to the precision of the eyetracker. Next we compute the offset fixation, by adding the offset distance in the offset angle to each true fixation location. We compute the capture rate as the percentage of offset fixations that fall within r degrees of the center of the object. To study the effect of stimulus size, we

vary x_{ul} and y_{ul}, and to study accuracy and precision, we vary the mean and standard deviation of the offset distance distribution. To study stimulus shape, we vary the proportion between x_{ul} and y_{ul}, thereby creating objects with a higher or a lower height-to-width ratio—that is, changing the ratio of perimeter to area. To study the effect of fixation distribution centrality, we draw the true fixation distribution from a beta distribution varying the alpha and beta parameters. The larger the beta-to-alpha parameter ratio, the more central the fixation distribution becomes. To study the offset angle, we draw offset angles uniformly between 0° and 360°, or if an offset angle tendency is assumed, we draw a single common offset angle from a uniform distribution between 0° and 360°.

The results of the simulation studies are shown in Fig. 1. The figure shows that larger stimulus sizes increase the capture rate, and that even for an excellent eyetracker, with accuracy = .5 and precision = .1, stimuli have to be more than 5° in diameter to achieve a high capture rate—that is, above .8. We also see that as accuracy and precision gradually decline, the capture rate goes down, but this is mostly true for small stimuli ≤2°, whereas large objects, ≥8°, retain a high capture rate even for very poor levels of accuracy and precision. We also see that the capture rate is influenced by the centrality of the fixation distribution, with more central distributions leading to higher capture rates. Finally, we see that as the area-to-perimeter ratio of a stimulus increases, the capture rate decreases and the variance of the capture rate increases. The ideal stimulus is therefore a circle, since it minimizes the area-to-perimeter ratio. Stimulus shapes such as rectangles are more vulnerable to offset angles, and therefore yield lower capture rates on average.

Generally, the simulations show that predicting the capture rate in a specific situation requires knowledge about the size and shape of the stimulus, the accuracy and precision of the eyetracker, and whether fixations are centrally distributed. We therefore recommend that studies that require high capture rates perform simulation studies beforehand. As an alternative to capture rate simulations, one can use a heuristic solution. If we assume that fixations are uniformly distributed and that our stimulus is circular, the capture rate can be approximated as the intersection between two displaced circles. This heuristic only holds when precision is very low, <.2, in which case the heuristic solution is identical to the actual one to the third decimal. To compute the heuristic, we only need to know

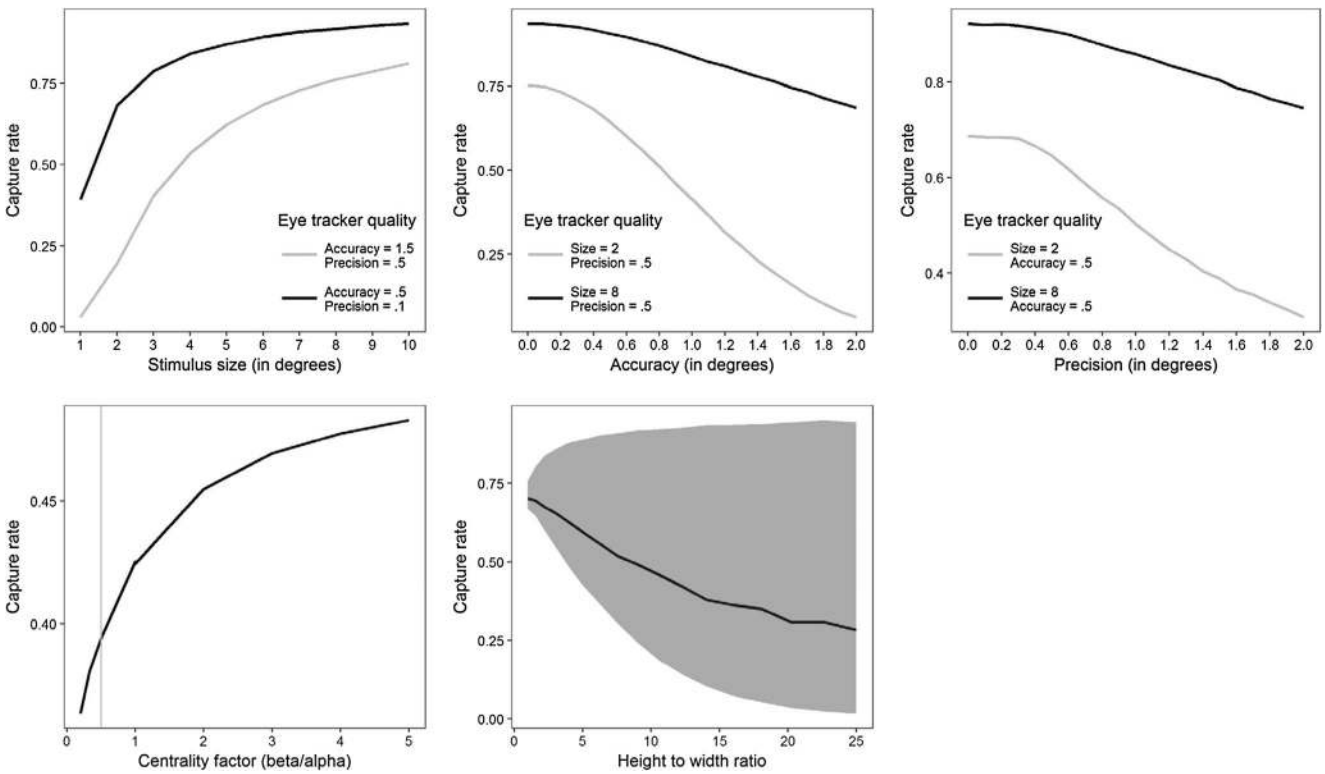


Fig. 1 Simulation results showing the expected capture rates depending on various factors. (Top left) Effects of stimulus size. (Top middle) Effects of eyetracker accuracy. (Top right) Effects of eyetracker precision. (Bottom left) Effects of the centrality of fixations to the object. The gray line indicates a uniform distribution of fixations, with more-central fixation distributions to the right of the line. (Bottom right) Effects of height-to-width ratio for rectangular stimuli. A ratio of 1 indicates a square, and higher ratios indicate more rectangular stimuli. The line shows the mean capture rate, and the shaded area indicates the minimum and maximum capture rates. The software for generating these calculations is available at https://github.com/jacoborquin/capturerate

the radius of the (round) stimulus, r, and the accuracy of the eyetracker, represented here as d:

$$\text{capture rate} = \frac{2r^2\cos^{-1}\left(\frac{d}{2r}\right) - \frac{1}{2}d\sqrt{4r^2 - d^2}}{\pi r^2}$$

When the precision of the eyetracker is 0, the heuristic solution is similar to the results obtained by simulation. It is important, however, that the heuristic be used only for round stimuli when we can safely assume uniform fixation distributions, and when the precision is below .2. In Table 1, we present the simulation results for six common eyetrackers, assuming round stimuli and a uniform fixation distribution.

Hidden defaults

A hidden default is a decision we are unaware of having made. Hidden defaults occur whenever we copy other researchers' experimental designs without considering alternatives, or when we analyze our eyetracking data unaware of the many transformations the software has performed on the data. The problem with hidden defaults is that they do not ensure an optimal result. In fact, hidden defaults are a guaranteed way of propagating poor ideas from researcher to researcher. As an example, many researchers may fail to realize that remote eyetrackers often average the positions of both eyes as a default, even though it is generally recommended to rely on the position of the dominant eye (Holmqvist et al., 2011, pp. 42, 60, 119). Of course, averaging might make sense in some situations. Both accuracy and precision have been found to improve when averaging the eyes (Cui & Hondzinski, 2006), but even with just a slight difference in timing between the two eyes, averaging the signals could alter saccade measures such as the latency, velocity profile, and peak velocity or skew. For studies in which these saccade measures are important, it is advisable to turn off averaging (Holmqvist et al., 2011, p. 60).

More generally, data processing in any eyetracker is largely a trade secret. Averaging can be turned off, but filtering is often hidden and can alter the saccade profile in ways that are very hard to remedy. Figure 2 shows how saccades have been given a very high onset acceleration, most likely by internal filtering.

Hidden defaults exist not only in software but also in specific lines of research. An example is the unfortunate use of high cutoffs for minimal fixation durations. For instance, Jansen, Nederkoorn, and Mulkens (2005) used a 300-ms minimum fixation duration threshold. Manor and Gordon (2003) noted that 200 ms has become the de facto standard in clinical studies, originally derived from a 1962 study of eye movements in reading. Since the range from 200 to 300 ms often encompasses the median of a fixation duration distribution (Holmqvist et al., 2011, p. 381), around 50% of the fixations will be lost with such a high cutoff, tending to change the results of a study entirely.

Less obvious hidden defaults only become evident with time. Saccade onset thresholds, hidden inside algorithms, guide how fast the eye must move before the movement can be considered a saccade. In a meta-analysis on Parkinson's disease, Chambers and Prescott (2010) surprisingly found that when tracking with video-based eyetrackers, patients have longer saccade latencies than controls, but not when tracked with scleral search coils (Robinson, 1963). They noted that Parkinson patients' saccades are subdued, meaning that the eye accelerates less vigorously. As a result, their saccades will typically take slightly longer to cross a saccade onset velocity threshold, even if the true latency is identical to that of controls. This effect is pronounced in video-based eyetracking, because the onset velocity threshold is higher than in the algorithms for coil data, which have less noise. In both cases, the saccade onset threshold is hidden in the software, inaccessible to the user. Saccade detection may work for control subjects and yet fail for clinical groups with nonnormal velocities. The only way to circumvent the problem of event detection is manual inspection, preferably of each saccade in each trial for each subject.

A simple remedy for hidden defaults is to map the flow of information and the data-processing steps, and to make active choices about each of these. Mapping the process, however, may be difficult, but help can be found in methodological overviews (Holmqvist et al., 2011; Schulte-Mecklenbeck et al., 2017).

Table 1 Minimum stimulus sizes, in degrees of visual angle, to obtain an 80% capture rate for a noncentral (uniform) fixation distribution, given the manufacturer-reported hardware accuracy and precision

Eyetracker	Accuracy	Precision	Min Size
EyeLink 1000 (ideal calibration)	.25	.01	1.6°
EyeLink 1000 (average calibration)	.5	.05	3.2°
Tobii 1750	.5	.25	3.3°
Tobii 2150	.5	.35	3.4°
SMI RED	.4	.03	2.6°
Eye Tribe	1	.1	6.4°

Total dwell time (also known as total gaze duration or total fixation duration)

The total dwell time (TDT) is the sum of all dwells (set of one or more consecutive fixations in an AOI) falling within an area of interest (AOI) during a trial or any other specified period of time (Holmqvist et al., 2011, pp. 190, 389). This metric is very popular and has been used in many published articles (Schulte-Mecklenbeck et al., 2017). The problem with TDT

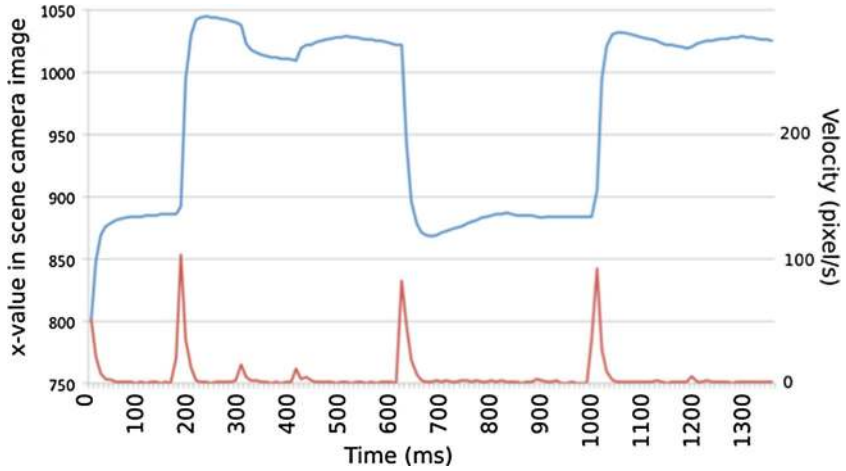


Fig. 2 Saccades recorded with the Tobii glasses II, 100 Hz. The red line is the velocity, and the blue line is the *x*-coordinate. The sharp onsets of saccades contrasts with smooth offsets, with no postsaccadic oscillations, suggesting that these saccade profiles are the result of a hidden filter. This suspicion is further supported by an RMS/STD value for this recording of 0.38, which is much lower than the expected 1.41 for unfiltered data (Holmqvist et al., 2017)

is that it often involves inappropriate aggregation of data. TDT becomes inappropriate when a researcher uses the metric to draw conclusions about one AOI receiving more attention than another AOI. Although it may be true that TDT is higher for AOI A than for AOI B, the difference in TDT can arise from three independent conditions. First, AOI A may receive more fixations or dwells than B; second, fixations to A may have a longer duration than fixations to B; and third, A may be fixated with a higher likelihood than B. Each of these three conditions has a different psychological interpretation.

- If A receives more dwells than B, even when both are fixated in all trials, this means that participants are more likely to refixate A. Refixations are probably due to top-down control, such as a high relevance of the stimulus to the task (Orquin & Mueller Loose, 2013) or the stimulus being confusing or difficult to process (Rayner, 2009).
- If the duration of fixations to A lasts longer than that of fixations to B, this can mean that A is the more complex stimulus, requiring a longer processing time (Just & Carpenter, 1976), or it may mean that A is the more interesting or relevant stimulus (Orquin & Mueller Loose, 2013).
- If A is more likely to be fixated than B, this could be due to both top-down and bottom-up control processes—that is, goal-driven versus stimuli-driven fixations. A bottom-up process would, for instance, imply that A is more salient than B, and therefore more likely to attract fixations (Itti & Koch, 2001). A top-down process would imply that A is more relevant than B, consequently attracting more fixations (Orquin & Lagerkvist, 2015).

Finding a difference in TDT only means that at least one of the three conditions has been met, and interpreting the difference requires breaking down the metric into its constituent parts.

To demonstrate this, we performed a reanalysis of the experiment reported in Orquin and Lagerkvist (2015). Their study investigated the effects of visual and motivational salience on eye movements in consumer choices. The study was a mixed within-subjects–between-subjects experiment in which participants made decisions between two food products, one of which bore a product label. The motivational salience of the label was manipulated between subjects by providing the participants with instructions about the label having a positive, a negative, or a neutral meaning. The visual salience of the label was manipulated within subjects as either high or low salience, by controlling the transparency of the label. We also analyzed the effect of product position. In the choice task, products were placed on the left or the right side of the screen, and we expected participants to have more eye movements to the left option in correspondence with their reading direction. To demonstrate the redundancy of TDT, we began by analyzing TDT and then proceeded to calculate fixation likelihood. Given a difference in fixation likelihoods, we analyzed fixation count, fixation duration, dwell count, and dwell duration conditionally on the AOI being fixated. We fitted all metrics with generalized linear mixed models by using the nlme package in R. To account for dependencies, we fitted random intercepts grouped by participant and trial.

The results of the analyses are shown in Table 2, and the observed effects are illustrated in Fig. 3. The left–right position of a product had a significant effect on TDT, with the left option having a higher TDT, as expected. Breaking down this effect, we found that there was no variance in the fixation likelihoods; all products were fixated in all trials. The difference in TDT therefore stems from one of the other metrics. In fact, all of the other metrics—fixation count, fixation duration,

Table 2 Significance tests for the breakdown of TDT in terms of its underlying metrics for three different factors: Position, plus visual and motivational salience

Dependent variable	Position	Visual Salience	Motivational Salience
Total dwell duration	$F(1, 1715) = 36.125$, $p < .001$	$F(1, 1044) = 4.897$, $p = .027$	$F(2, 147) = 1.512$, $p = .224$
Fixation likelihood	No variance in fixation likelihood	$F(1, 1044) = 8.205$, $p = .004$	$F(2, 147) = 11.79$, $p < .001$
Fixation count	$F(1, 1715) = 36.298$, $p < .001$	$F(1, 567) = 2.514$, $p = .113$	$F(2, 141) = 0.008$, $p = .992$
Fixation duration	$F(1, 1715) = 12.669$, $p < .001$	$F(1, 567) = 0.892$, $p = .345$	$F(2, 141) = 2.57$, $p = .080$
Dwell count	$F(1, 1715) = 574.495$, $p < .001$	$F(1, 567) = 0.244$, $p = .622$	$F(2, 141) = 2.498$, $p = .086$
Dwell duration	$F(1, 1715) = 27.673$, $p < .001$	$F(1, 567) = 0.522$, $p = .470$	$F(2, 141) = 1.448$, $p = .238$

dwell count, and dwell duration—were significantly different. The left option received more fixations and dwells, but the right option had longer fixations and dwells. Visual salience had a marginally significant effect on TDT, and this effect was explained entirely by differences in fixation likelihood, with the high-salience label being more likely to be fixated than the low-salience one. Given that the label was fixated, there were no differences in any of the other metrics. Motivational salience had no effect on TDT, but our breakdown approach revealed that there was nevertheless a significant difference in fixation likelihood, as well as marginal effects on fixation duration and dwell count. We concluded from this reanalysis that given a difference in TDTs, we cannot know what underlying metric drives this difference. Given that no difference in TDTs is present, we also cannot conclude that there are also no differences in the underlying metrics. For this reason, we advise against the use of TDT in eyetracking research.

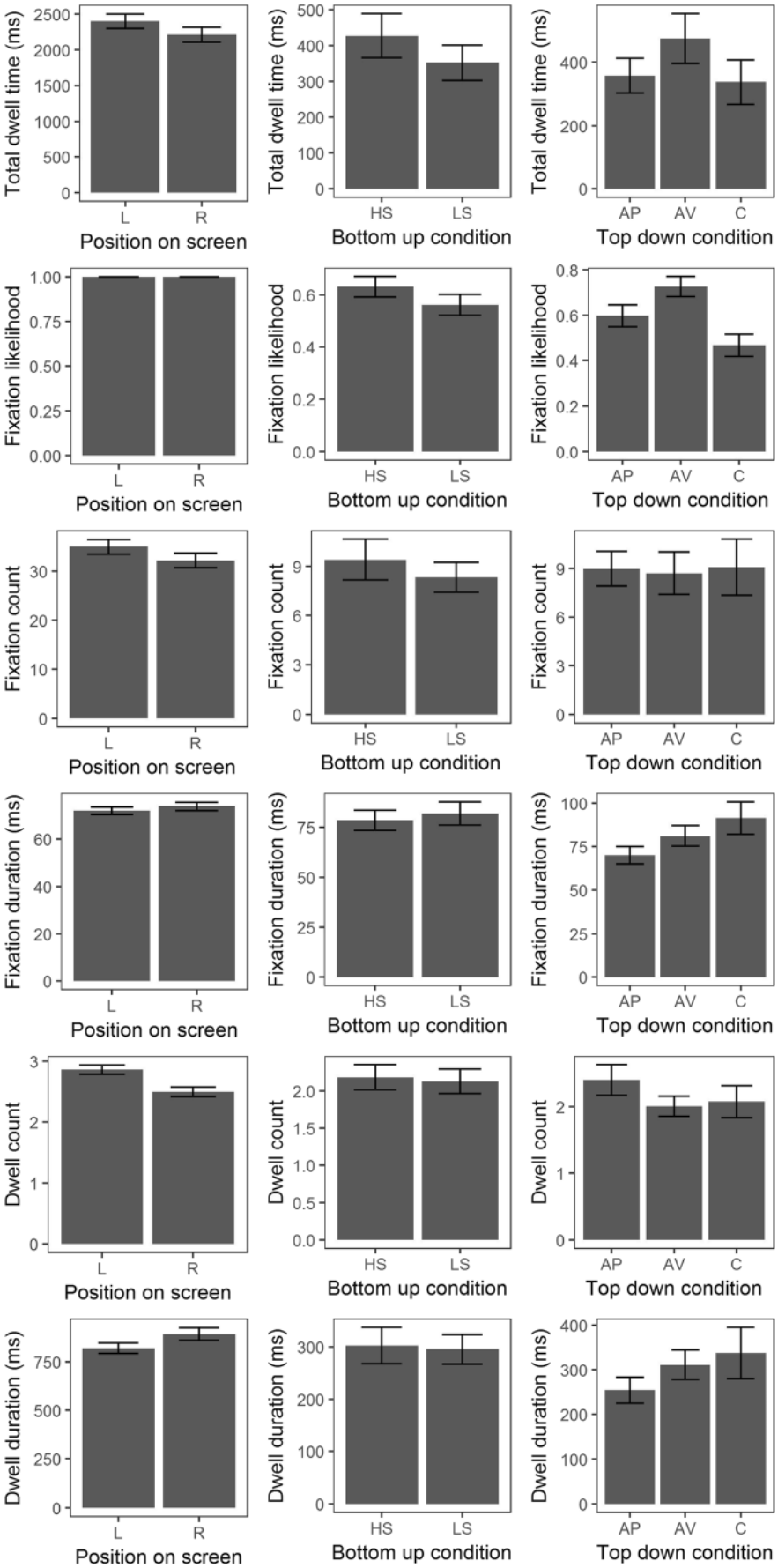
Fixed versus free exposure time

When designing eyetracking experiments, we must decide on the duration of stimulus exposure. A common approach is to fix the exposure time so that a participant sees a stimulus for some predetermined period of time (Reutskaja, Nagel, Camerer, & Rangel, 2011). The alternative, using a free exposure time, allows participants to gaze at the stimulus for as long as they wish, typically until the participant presses a key on the mouse or the keyboard. Although a fixed exposure time has its merits in, for instance, psychophysics, it tends to be misapplied in more behavior-oriented studies. The problem is twofold. First, it is difficult to match the exposure time to the exact point in time at which the participant would have otherwise terminated the trial. A fixed exposure times will therefore always be either shorter or longer than the participant-driven exposure time. This deviation will most likely create an experience of either *time pressure* (Reutskaja et al., 2011) or *idleness* (Hsee, Yang, & Wang, 2010). In many cases, time pressure is what the experimenter hopes to achieve—idleness probably is not. The second problem with a fixed exposure time is interpretation of the data. Assuming idleness, one must consider the distribution of eye movements in the idle period. For example, in a discrete-choice experiment with fixed exposure time, one has a clear interpretation of eye movements until the decision is made. In the idle period, however, the participant may stare at any object at random or continue in a postdecision process (Clement, 2007). As a rule, it is therefore advisable not to use a fixed exposure time, but there are, of course, situations in which it is required. If we, for instance, wish to understand the development of a fixation process over time, a fixed exposure time allows for direct comparison of different trials. Using free exposure times, on the other hand, requires that we transform trials of different lengths or focus our analysis on, for instance, the first 500 ms after stimulus onset or the last 500 ms before a response is made (Shimojo, Simion, Shimojo, & Scheier, 2003).

Assuming an eye–mind relationship (reverse inference)

It can be very tempting to think that eyetrackers report attention or some other cognitive process. Eyetrackers, however, report eye movements and gaze, while attention is always inferred. Nevertheless, because attention plays a central part in many models of cognition, researchers often assert the so-called *eye–mind assumption*, which was proposed by Just and Carpenter (1976). On the basis of studies of eye movements in reading, they suggested that there is no appreciable lag between what is being fixated and what is being processed at a cognitive level. The eye–mind assumption originated from reading research but has been introduced into other areas, as well (Svenson, 1979).

There is, indeed, a relation between looking and thinking, but this relation must be proved rather than just assumed, because of its many caveats and exceptions. For instance, eye movements are closely coupled with attention, such that a saccade is always preceded by a change in attention (Deubel & Schneider, 1996). However, because attention shifts before the fixation ends, attention and fixations are not perfectly coupled. In fact, the eye–mind assumption has been falsified in various instances. For instance, Deubel (2008) has shown disassociations of fixations and attention by up to 250 ms in


Total dwell time (ms)
Position on screen
L
R
Bottom up condition
HS
LS
Top down condition
AP
AV
C
Fixation likelihood
Fixation count
Fixation duration (ms)
Dwell count
Dwell duration (ms)

◀ **Fig. 3** Effects of position on screen, visual salience (bottom-up condition), and motivational salience (top-down condition) on total dwell time and its underlying metrics

some situations. For all these reasons, the eye–mind assumption should only be made after careful deliberation.

Instead of the eye–mind assumption, which is difficult to support, eyetracking researchers may instead consider a *signal detection assumption*. The question is whether fixations to an object imply that the object has been processed, and whether the absence of fixations implies that the object has not been processed. We can then consider situations that lead to false positives (fixated but not processed) and false negatives (not fixated but processed).

One of the situations that may lead to false negatives is the possibility of peripheral processing—that is, an observer detecting and identifying an object without fixating it. The influence of peripheral vision is well established in both reading and scene viewing (Rayner, 2009), and peripherally processed words can lead to semantic activation and priming effects (Devine, 1989). One of the challenges in ruling out peripheral uptake is that it depends on the features of the stimuli, such as the size and contrast of objects (Melmoth & Rovamo, 2003) or how crowded the scene is around the object (Whitney & Levi, 2011), as well as on characteristics of the observer, such as the level of expertise and familiarity with the task (Reingold, Charness, Pomplun, & Stampe, 2001).

One of the situations that may lead to false positives is the risk of selective feature extraction. It has been demonstrated that observers typically fail to extract or encode all possible features from visual objects, only extracting or encoding the task-relevant features (Hayhoe, Bensinger, & Ballard, 1998). This means that we cannot conclude from a fixation to an object that the object as a whole has been processed. Instead, the observer may only have processed a single feature of the object. A related phenomenon is *inattentional blindness*, in which observers make a direct fixation to an object yet are unaware of the existence of the fixated object (Koivisto, Hyönä, & Revonsuo, 2004).

Another issue that may lead to both false positives and false negatives is inappropriate AOI definitions. Because of inaccuracies in both eyetrackers and the human visual system, fixations often fall outside the object that is the target of the saccade. If the AOI around an object has a narrow margin—for example, <0.5° beyond the object border—we may fail to detect fixations falling outside the object, leading to false negatives. On the other hand, when objects are placed close to each other, we risk assigning fixations that fall outside an object to a neighboring object, leading to false positives for the neighboring object (Orquin et al., 2016).

Finally, it is worth mentioning that other data sources—for example, choice data, verbal protocols, and retention tests—can suggest whether the object was processed and taken into consideration. This is known as *methodological triangulation* (Holmqvist et al., 2011, p. 95).

Threats to external validity

Undersampling of naturalistic stimuli

As we discussed above, it is regrettably common to find eyetracking studies with only one or two critical trials (Schulte-Mecklenbeck et al., 2017). Besides the fact that a limited number of trials leads to lower statistical power, it leads to another negative consequence. Whenever studies rely on naturalistic stimuli—for instance, images of products or advertising—one necessarily factors into the experiment any random features of those stimuli. Some images may be more or less bright, include more or larger objects, and so forth. Eye movements are highly susceptible to these stimulus differences (Orquin & Mueller Loose, 2013). However, these differences are not a problem as such. We can think of the experimental stimuli as a random effect; in this case, the more trials we include, the safer it is to assume that any differences wash out over the conditions of interest. Including more and heterogeneous stimuli, then, actually adds to the robustness of the conclusions (Cooper, Hedges, & Valentine, 2009). Experiments with only one or two trials, on the other hand, produce eye-movement distributions that are specific to the particular stimulus. As a rule of thumb, using more trials always reduces the bias in our stimulus sample. We can calculate the expected deviation, E[d], of a sample of size N from a normally distributed population as:

$$\mathrm{E[d]} = \int_0^\infty 2xf(x)dx$$

where $f(x)$ is the probability density function for a normal distribution with mean equal zero, and standard deviation, $\sigma = s/\sqrt{N}$, where s is the population standard deviation. As N increases, the standard deviation of the sample, σ, decreases and the expectation of the sample going toward zero, which is the population mean. Following this we see that having more than 16 trials yield an average bias <0.2 *SD*—that is, a small effect in terms of Cohen's *d*. Using one trial yields an average bias >0.75 *SD*—that is, a large effect, meaning that our sample is biased or unrepresentative of the population. If we assume that the stimuli differ on more features—for example, visual salience, surface size, and position—the probability of at least one feature being biased is $1 - P^k$, where P is the probability of the feature being biased, and k is the number of features. To demonstrate the importance of adequate sampling, we reanalyzed data from Peschel and Orquin (2013). Their data set was based on a list of 158 consumer products from four categories sold in Danish supermarkets. The product features—for example, brand, logo, image,

and nutrition labels—were described with regard to their visual salience, relative surface size, and distance to the center of the product, dimensions known to influence the probability of consumers fixating nutrition labels (Graham, Orquin, & Visschers, 2012). Our question was, how many products should we include in a study in order to reliably estimate the probability of consumers fixating nutrition labels? If we only include one product, we are likely to either over- or underestimate the probability of consumers fixating the label by a large margin. To understand how many products we would need for a representative sample, we focused on the 80 products that carried nutrition labels. We drew sample sizes from 1 to 25 products. For each sample size, we iterated 10,000 times and computed the absolute deviation of the sample mean from the population mean. We then divided by the population standard deviation to obtain a standardized effect size measure: $|M_{\text{sample}} - M_{\text{population}}|/SD_{\text{population}}$. The results of the simulation are shown in Fig. 4. The figure is nearly identical to the analytical solution, showing that a representative sample, defined as deviating by less than 0.2 *SD*s from the population on all three dimensions, on average requires 16 products.

Generalization of eye-movement distributions

Applied research often wishes to make inferences about classes of stimuli such as advertising, product packaging, health warnings, and so forth, for policy purposes (Graham et al., 2012). If the experiment suffers from undersampling of naturalistic stimuli, then clearly we cannot generalize anything beyond the sparse stimuli. Even if the experiment uses a broad range of stimuli, it may still be difficult to generalize eye movements beyond the laboratory environment. As we discussed above, eye movements are highly susceptible to small changes in the environment. In a laboratory setting, we may find that participants exposed to faces fixate directly on the eyes. Generalizing this eye-movement distribution to the real world would, however, be problematic, since people in natural environments mostly fixate just below the eyes (Foulsham, Walker, & Kingstone, 2011).

One remedy of this problem would be to change the focus from eye-movement distributions to psychological mechanisms. A causal mechanism is our best chance of generalizing beyond the laboratory (Cooper et al., 2009). For instance, a psychological mechanism such as central gaze bias—that is, a

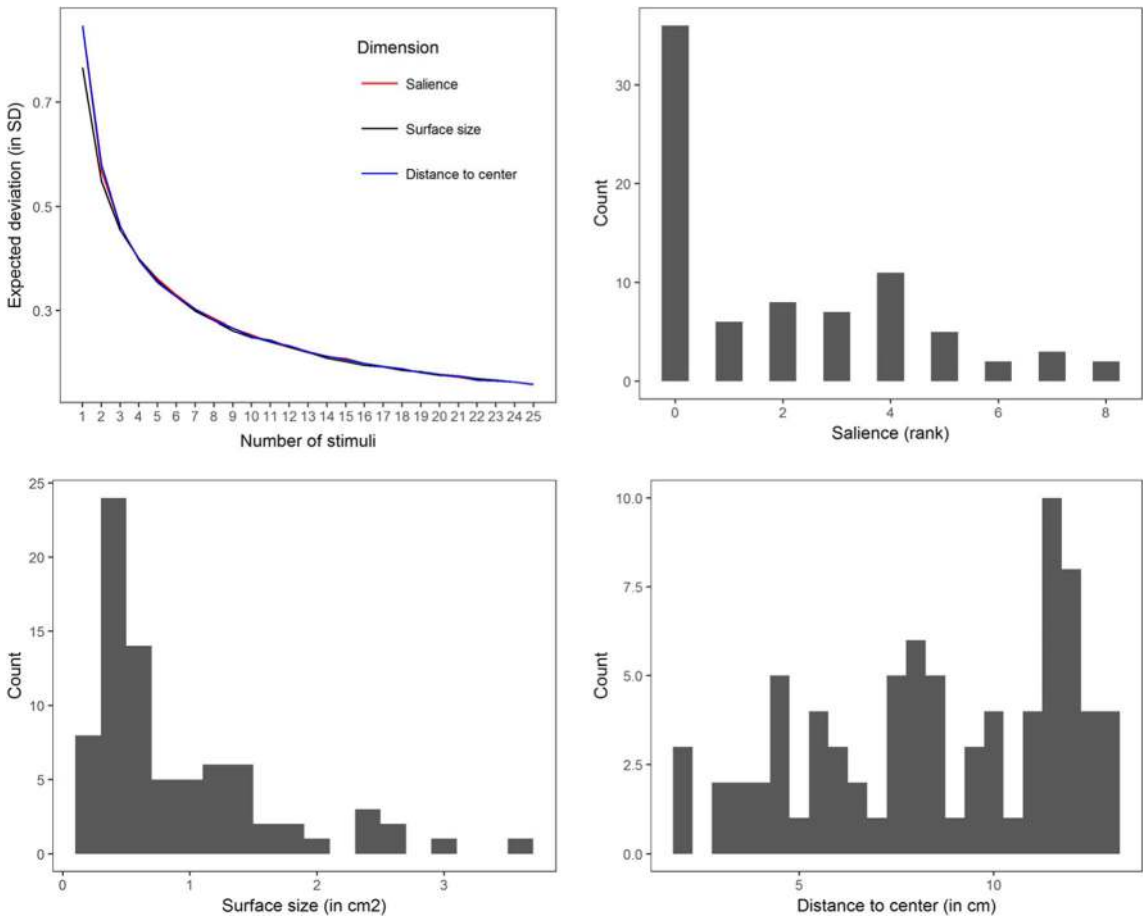


Fig. 4 (Top left) Expected deviations (as standard deviations [*SD*]) between a stimulus sample of size *N* and the stimulus population of nutrition labels, (Top right) Histogram of salience ranks. (Bottom left) Histogram of surface size. (Bottom right) Histogram of distance to center

tendency to fixate the center of an array of products—may transfer well from the laboratory to the supermarket (Gidlöf & Holmqvist, 2011). Mechanism studies, however, impose greater demands on the research question and experimental design. First, we need to identify possible mechanisms based on known or new theoretical considerations about eye-movement control processes. Second, on the basis of the specific hypothesis, we need a true experimental design with random assignment to treatment conditions; that is, besides our manipulation of the independent variable, everything else has to remain equal. Using a quasi-experimental design, Lohse (1997), for example, studied the effect of surface size on eye movements to yellow-page advertising. Even though the study was informative about the effect of surface size, in theory it is impossible to make causal claims about surface size, because it could be confounded with other variables. Third, given that we hypothesized a causal mechanism, conducted a true experiment, and established a statistical effect on eye movements, we would still have to exercise caution in making any claims about causality. Only in the absence of alternative explanations and successful replications of our hypothesis could we have confidence in the causal mechanism.

Summary

Eyetracking research has experienced a surge in the past decade as the equipment has become cheaper and easier to use. Many types of eyetrackers can be operated without any skills in experimental design or data analysis, thereby lowering the barriers to conducting eyetracking research. This development may have led to some research practices that would best be avoided. Motivated by this concern, we have proposed a list of threats to the validity of eye-movement research. The list of threats will allow researchers to identify problems before conducting their studies and may serve as a reference for editors and reviewers. It is important, however, to realize that this list cannot replace what has already been said about sound research practices, and that the list may not be exhaustive. New threats may be added as methodological research progresses. Also, we must emphasize that the list should never be applied uncritically, lest it become a hidden default.

Author note The authors thank Ignace Hooge, Richard Dewhurst, and Sonja Perkovic for comments on previous versions of the manuscript.

References

Antúnez, L., Vidal, L., Sapolinski, A., Giménez, A., Maiche, A., & Ares, G. (2013). How do design features influence consumer attention when looking for nutritional information on food labels? Results from an eye-tracking study on pan bread labels. *International Journal of Food Sciences and Nutrition*, *64*, 515–527. https://doi.org/10.3109/09637486.2012.759187

Baker, D. A., Schweitzer, N. J., Risko, E. F., Ware, J. M., & Sinnott-Armstrong, W. (2013). Visual attention and the neuroimage bias. *PLoS ONE*, *8*, e74449. https://doi.org/10.1371/journal.pone.0074449

Chambers, J. M., & Prescott, T. J. (2010). Response times for visually guided saccades in persons with Parkinson's disease: A meta-analytic review. *Neuropsychologia*, *48*, 887–899. https://doi.org/10.1016/j.neuropsychologia.2009.11.006

Clement, J. (2007). Visual influence on in-store buying decisions: An eye-track experiment on the visual influence of packaging design. *Journal of Marketing Management*, *23*, 917–928. https://doi.org/10.1362/026725707X250395

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). The handbook of research synthesis and meta-analysis (2nd). New York: Russell Sage Foundation.

Cui, Y., & Hondzinski, J. M. (2006). Gaze tracking accuracy in humans: Two eyes are better than one. *Neuroscience Letters*, *396*, 257–262. https://doi.org/10.1016/j.neulet.2005.11.071

Deubel, H. (2008). The time course of presaccadic attention shifts. *Psychological Research*, *72*, 630–640. https://doi.org/10.1007/s00426-008-0165-3

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*, 1827–1837. https://doi.org/10.1016/0042-6989(95)00294-4

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Dodd, M. D., Balzer, A., Jacobs, C. M., Gruszczynski, M. W., Smith, K. B., & Hibbing, J. R. (2012). The political left rolls with the good and the political right confronts the bad: Connecting physiology and cognition to preferences. *Philosophical Transactions of the Royal Society B*, *367*, 640–649. https://doi.org/10.1098/rstb.2011.0268

Donovan, T., & Litchfield, D. (2013). Looking for cancer: Expertise related differences in searching and decision making. *Applied Cognitive Psychology*, *27*, 43–49. https://doi.org/10.1002/acp.2869

Duchowski, A. T. (2007). Eye tracking methodology: Theory and practice (2nd). New York: Springer Science & Business Media.

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*, 1920–1931. https://doi.org/10.1016/j.visres.2011.07.002

Gidlöf, K., & Holmqvist, K. (2011). *Expansion of the central bias, from computer screen to the supermarket*. Paper presented at the 16th European Conference on Eye Movements (ECEM, 2011), Marseille, France. Abstracted in *Journal of Eye Movement Research*, *4*, 260.

Glöckner, A., & Herbold, A.-K. (2011). An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making*, *24*, 71–98. https://doi.org/10.1002/bdm.684

Graham, D. J., Orquin, J. L., & Visschers, V. H. M. (2012). Eye tracking and nutrition label use: A review of the literature and recommendations for label enhancement. *Food Policy*, *37*, 378–382. https://doi.org/10.1016/j.foodpol.2012.03.004

Hartridge, H., & Thomson, L. C. (1948). Methods of investigating eye movements. *British Journal of Ophthalmology*, *32*, 581–591. Retrieved from www.ncbi.nlm.nih.gov/pubmed/18170495

Hayhoe, M. M., Bensinger, D. G., & Ballard, D. H. (1998). Task constraints in visual working memory. *Vision Research*, *38*, 125–137. https://doi.org/10.1016/S0042-6989(97)00116-8

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Halszka, J., & van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford: Oxford University Press.

Holmqvist, K., Zemblys, R., Cleveland, D., Mulvey, F., Borah, B., & Pelz, J. (2015). The effect of sample selection methods on data quality measures and on predictors for data quality. Paper presented at the European Conference on Eye Movements, Vienna.

Holmqvist, K., Zemblys, R., Niehorster, D. C., & Beelders, T. (2017). Magnitude and nature of variability in eye-tracking data. In Proceedings of the 19th European Conference on Eye Movements. ECEM: Wuppertal

Hsee, C. K., Yang, A. X., & Wang, L. (2010). Idleness aversion and the need for justifiable busyness. *Psychological Science*, *21*, 926–930. https://doi.org/10.1177/0956797610374738

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*, 194–203. https://doi.org/10.1038/35058500

Jansen, A., Nederkoorn, C., & Mulkens, S. (2005). Selective visual attention for ugly and beautiful body parts in eating disorders. *Behaviour Research and Therapy*, *43*, 183–196. https://doi.org/10.1016/j.brat.2004.01.003

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*, 441–480. doi:https://doi.org/10.1016/0010-0285(76)90015-3

Koivisto, M., Hyönä, J., & Revonsuo, A. (2004). The effects of eye movements, spatial attention, and stimulus features on inattentional blindness. *Vision Research*, *44*, 3211–3221. https://doi.org/10.1016/j.visres.2004.07.026

Lohse, G. L. (1997). Consumer eye movement patterns on yellow pages advertising. *Journal of Advertising*, *26*, 61–73. https://doi.org/10.1080/00913367.1997.10673518

Manor, B. R., & Gordon, E. (2003). Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of Neuroscience Methods*, *128*, 85–93. https://doi.org/10.1016/S0165-0270(03)00151-1

Melmoth, D. R., & Rovamo, J. M. (2003). Scaling of letter size and contrast equalises perception across eccentricities and set sizes. *Vision Research*, *43*, 769–777. https://doi.org/10.1016/S0042-6989(02)00685-5

Monty, R. A. (1975). An advanced eye-movement measuring and recording system. *American Psychologist*, *30*, 331–335. https://doi.org/10.1037/0003-066X.30.3.331

Niehorster, D. C., Cornelissen, T. H. W., Holmqvist, K., Hooge, I. T. C., & Hessels, R. S. (2017). What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-017-0863-0

Nummenmaa, L., Hyönä, J., & Calvo, M. G. (2006). Eye movement assessment of selective attentional capture by emotional pictures. *Emotion*, *6*, 257–268. https://doi.org/10.1037/1528-3542.6.2.257

Orquin, J. L., Ashby, N. J. S., & Clarke, A. D. F. (2016). Areas of interest as a signal detection problem in behavioral eye-tracking research. *Journal of Behavioral Decision Making*, *29*, 103–115. https://doi.org/10.1002/bdm.1867

Orquin, J. L., & Lagerkvist, C. J. (2015). Effects of salience are both short- and long-lived. *Acta Psychologica*, *160*, 69–76. https://doi.org/10.1016/j.actpsy.2015.07.001

Orquin, J. L., & Mueller Loose, S. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, *144*, 190–206. https://doi.org/10.1016/j.actpsy.2013.06.003

Peschel, A. O., & Orquin, J. L. (2013). A review of the findings and theories on surface size effects on visual attention. *Frontiers in Psychology*, *4*, 21–30. https://doi.org/10.3389/fpsyg.2013.00902

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*, 1457–1506. https://doi.org/10.1080/17470210902816461

Reingold, E. M. (2014). Eye tracking research and technology: Towards objective measurement of data quality. *Visual Cognition*, *22*, 635–652. https://doi.org/10.1080/13506285.2013.876481

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science*, *12*, 48–55. https://doi.org/10.1111/1467-9280.00309

Reutskaja, E., Nagel, R., Camerer, C. F., & Rangel, A. (2011). Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review*, *101*, 900–926. https://doi.org/10.1257/aer.101.2.900

Robinson, D. (1963). A method of measuring eye movemnent using a scieral search coil in a magnetic field. *IEEE Transactions on Bio-Medical Electronics*, *10*, 137–145. https://doi.org/10.1109/TBMEL.1963.4322822

Russo, J. E. (2011). Eye fixations as a process trace. In M. Schulte-Mecklenbeck, A. Kühberger, J. G. Johnson, & R. Ranyard (Eds.), A handbook of process tracing methods for decision research: A critical review and user's guide (pp. 43–64). New York: Psychology Press.

Schulte-Mecklenbeck, M., Fiedler, S., Renkewitz, F., & Orquin, J. L. (2017). Reporting standards in eye-tracking research. In M. Schulte-Mecklenbeck, A. Kühberger, & J. Johnson (Eds.), A handbook of process tracing methods. New York: Routledge.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference (Technical Report). Wadsworth Cengage Learning, Independence. Retrieved from https://pdfs.semanticscholar.org/f141/aeffd3afcb0e76d5126bec9ee860336bee13.pdf

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, *14*, 47–69. https://doi.org/10.1023/A:1013180410169

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, *6*, 1317–1322. https://doi.org/10.1038/nn1150

Svenson, O. (1979). Process descriptions of decision making. *Organizational Behavior and Human Performance*, *23*, 86–112. https://doi.org/10.1016/0030-5073(79)90048-5

Towal, R. B., Mormann, M., & Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, E3858–E3867. https://doi.org/10.1073/pnas.1304429110

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, *94*, 119–133. https://doi.org/10.1016/j.jml.2016.10.003

Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4):160–168. https://doi.org/10.1016/j.tics.2011.02.005

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832