Journal of Eye Movement Research
3(3), 6, 1-12

Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more

Richard Andersson
Lund University Cognitive Science

Marcus Nyström
Lund University Humanities Lab

Kenneth Holmqvist
Lund University Humanities Lab

We use simulations to investigate the effect of sampling frequency on common dependent variables in eye-tracking. We identify two large groups of measures that behave differently, but consistently. The effect of sampling frequency on these two groups of measures are explored and simulations are performed to estimate how much data are required to overcome the uncertainty of a limited sampling frequency. Both simulated and real data are used to estimate the temporal uncertainty of data produced by low sampling frequencies. The aim is to provide easy-to-use heuristics for researchers using eye-tracking. For example, we show how to compensate the uncertainty of a low sampling frequency with more data and post-experiment adjustments of measures. These findings have implications primarily for researchers using naturalistic setups where sampling frequencies typically are low.

Keywords: eye-tracking measures, simulation, temporal sampling frequency, reliability, fixation duration, saccadic latency, one-point measures, two-point measures, temporal sampling error, validity

Introduction

The role of sampling frequency and its mathematical implications to the resulting data may be common knowledge to statisticians, but in the eye-tracking community, sampling frequency is rarely highlighted in the methodological discussions. Sampling frequency is not the only important property of eye-trackers (precision, accuracy and versatility on participants are others), but it is definitely the most, by manufacturers, highlighted technical property. Manufacturers use sampling frequency as a major sales argument, and sampling frequency is the most mentioned technical property of eye-trackers in journal papers. There are good reasons for this: Sampling frequency affects what you can do with your eye-tracker in a number of ways. Some uses, such as precise measurements of small saccades, are so sensitive that a higher sampling frequency (and precision) is necessary to estimate them accurately (see e.g. the discussion in the supplemental methods of Kagan, Gur, & Snodderly, 2008). However, there are also cases where a naturalistic setup demands eye-trackers that currently lack the speed of their stationary counterparts. Consider the following quote from Green (2002):

> Similarly, Crundall & Underwood (1998) reported that experienced drivers had shorter fixation durations for suburban roads (324 versus 335 ms) and divided highways (349 versus 395 ms), but not for rural roads (381 versus 364 ms). Although statistically significant differences are claimed, these differences are at the limits of recording accuracy at 30 Hz (33 milliseconds per video frame).

Sampling frequency is measured using the unit Hertz (Hz), which refers to the number of samples per second. Most modern eye-trackers have sampling frequencies ranging from 25 - 2000 Hz. For the many 50 Hz eye-trackers, a sample is registered once every 20 ms, whereas a 250 Hz eye-tracker samples every 4 ms. We seem to think that faster is automatically better, like we think more pixels in a digital camera is always better, but it is also reasonable to expect the marginal benefit for every further Hz to diminish at some level. For instance, the benefit of a 2000 Hz system over a 1000 Hz system should not equal that of a 100 Hz over a 50 Hz eye-tracker, even though both constitute a doubling of the frequency. It is currently not clear what sampling

The authors wish to thank Kazuo Koga of EcoTopia Science Institute, Nagoya University, Japan, and one anonymous reviewer for extremely helpful reviews. The authors also thank the eye-tracking group at Lund University Humanities Lab for valuable comments. Source code for the simulations can be found at the main author's webpage: http://www.humlab.lu.se/richard

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

frequency is necessary for what effect size, and standards vary.

This points to the question: what sampling frequency and/or data amount is necessary to be certain of eye-tracking results where the sampling-related uncertainty exceeds the effect magnitudes found?

For oscillating eye-movements, such as tremors, we can argue based on the Nyquist-Shannon sampling theorem (Shannon, 1949) that the sampling frequency should be at least twice the speed of the particular eye movement (e.g., behaviour at 150 Hz requires >300 Hz sampling frequency). Other than that, the typical practice is to use whatever is used in your particular field of research or faster. Low-level visual cognition research can use constraining setups favouring systems with speeds from 1000 Hz to 2000 Hz, as naturalism is not typically a primary concern, but rather to maintain control over the variables. Research using gaze-contingent display changes, for example in real-time exchanging peripheral letters with 'x' to manipulate parafoveal preview benefits in a reading task (e.g., McConkie & Rayner, 1975), are usually the most demanding experiments in terms of frequency. This is because high speeds allow the system to detect saccade launches earlier and provide the display changes even faster, which minimizes the risk of the participant noticing the manipulation. Research investigating higher-level cognition and using naturalistic tasks commonly prefer systems allowing free movement of the head, either by remote eye-tracking or head-mounted and mobile eye-tracking. These systems typically operate at speeds from 25 Hz to 250 Hz. A specific community of researchers choose to use web-cameras as eye-trackers, with the goal of bringing inexpensive gaze-interfacing capabilities to the masses, and these cameras typically have sampling frequencies below or equal to 30 Hz. Also, analyses using video are typically limited the frame speed, which most often is around 24 fps/Hz. But even high-end eye-tracking systems allow different setups that exchange sampling frequency for binocularity or remote filming. This requires us to know what speeds we need for our particular research questions and also to know when it yields a net improvement to sacrifice speed in order to capture the behaviour in a more naturalistic setting, e.g., using the less intruding remote filming for slower eye movements.

Additionally, sampling frequency heavily affects many measures we use, and what we can use them for. For instance, Enright (1998) provides evidence that saccadic peak velocity can be well estimated in 60 Hz data from eye-trackers using the relation between pupil and corneal reflection, but only for saccades larger than 10°. For saccades shorter that 10°, typical of reading, the peak velocity calculation is not accurate with 60 Hz data. Juhola, Jäntti, and Pyykkö (1985), using Electrooculography (EOG) and photoelectric eye-trackers to study 20° saccades, argue that sampling frequency should be higher than 300 Hz to accurately calculate the maximum saccadic velocity. Inchingolo and Spanio (1985) found that saccadic duration and velocity data from a 200 Hz EOG system that they tested are equivalent to the same data from a 1000 Hz system, but only for saccades larger than 5°.

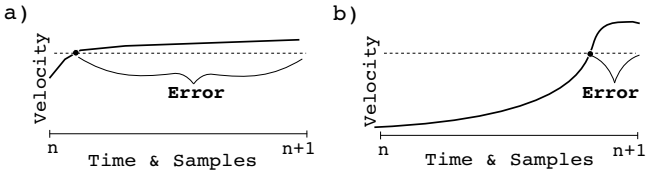

Figure 1. Consider the measurement of the event that triggers a saccade velocity criterion (dashed line). The temporal sampling error occurs when the eye increases velocity after a recent sampling of the eye, n, resulting in changes not being registered until the next sample $n+1$. Case a) shows a large error when the eye accelerates right after it just having been sampled, resulting in an error equal to almost the duration of one sample. Case b) shows a small error, where the eye accelerates at the end of the current sample and just before the next sample.

Whereas previous studies have focused on sampling frequency and its role for particular, often saccade-related, measures, we explore the effect of sampling-related errors more generally. Our aim is to explain the source of these errors, describe them mathematically, simulate their effects in actual experiments and provide easy-to-follow heuristics to compensate for these effects.

The source of the error

As it is impossible to have an infinite sampling frequency, each eye-tracker instead takes an instantaneous snapshot of the eye at a fixed rate (typically 25-2000 Hz). Each snapshot is a point in time, taken to be representative of a whole interval of time. For instance, with a 50 Hz system, the position of the eye at each sample is assumed to be valid for the whole 20 ms, even if it is very likely that the eye did not have that exact position just before the moment of sampling. The eye-tracker cannot sample the eye in a position which it has not moved to yet, but the system may sample the eye in the correct position or a position it recently had. By necessity, sampling always lags behind the position of the eye because the eye is constantly moving to some extent. Figure illustrates the resulting temporal sampling error, where the eye-tracker mis-estimates the correct point in time that a particular event (the triggering of a velocity criterion) takes place.

It should be noted that this paper addresses temporal measures, i.e., any measures that tries to estimate either the duration of an event or the point in time when an event takes place. These temporal measures are estimated using two reference points in time. The two

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

points we will call the *start criterion* and the *stop criterion*. The exact operationalization of the criteria will vary with the specific measures, but typically we focus on the system clock time of these events as they occur. For example, if an event starts at time 1 and it stops at time 10, then the resulting duration of the event is $10-1=9$ in whatever time unit we are measuring in. If we want to estimate a point in time when something occurs rather than a duration, we typically set the start event criterion to be the point when we start counting the time from 0. For example, if we want to estimate the point in time when the eye makes a particular movement in a trial, then typically we start counting the time from the beginning of the trial (our zero point). In this case, the trial start is our start criterion.

Throughout this paper, we will use the term *sampling point* to refer to that particular point in time when the eye image is captured by the eye-tracker. We use the term *sample* to refer to the eye image and the resulting coordinate pair that is taken to be valid for a period of time related to the sampling frequency of the eye-tracker. We use the term *window of no sampling* to refer to the time that passes between the two sampling points. The term *temporal sampling error*, or simply error, will refer to the time between the point of actual objective occurrence of an event and the detected occurrence of an event, e.g., the time between the point where the gaze enters an area of interest and the point the system actually registers the gaze inside the area. What is referred to as a temporal sampling error in this paper seems to be same phenomenon that Kagan et al. (2008) calls "temporal offset´´ (in the caption of Supplemental Figure 6). Do not confuse this temporal sampling error with other errors of measurements, such as spatial offset. The distribution of means of temporal sampling errors will often be modulated by the amount of data we have, and we will use the term *data points* to refer to the number (count) of a particular measure we have recorded, for example the number of fixation durations, dwell times, saccade durations and so on.

We make the very plausible assumption that a true oculomotor event, e.g. the eye passing a $60^\circ/s$ velocity criterion of saccade detection, is equally likely to occur anytime between two sampling points (i.e., uniformly distributed). For this paper, we also assume an eye-tracking system with zero system latency and cameras which take snapshots of the eye rather than continuously transmitting camera pixels going from the top-left corner to the bottom-right corner.

The one-point temporal sampling error

Consider a visual search task where we are interested in how fast participants can locate a target object in a cluttered scene. We may use, as a dependent variable, a saccadic latency measure where we measure the duration from the onset of a stimulus until a saccade is launched towards a designated target on the screen ($duration = time_{stop} - time_{start}$). If the time-stamp of the stimulus onset is 5674 ms and the saccade to the target is detected to be launched at 6743 ms, then the resulting saccadic latency is $6743-5674=1069$ ms. This precise saccadic latency value, however, assumes that we correctly detected the saccade launch at exactly 6743 ms. We will now describe how a temporal sampling error occurs for this example. Assuming no system latency, the onset of the stimulus will appear at the same time as the system timer starts counting the designated trial durations. The control computer records the processed eye-images from the onset of the stimulus until the offset of the same. During our analysis, we extract only the eye data corresponding to the time between the stimulus onset and until the first saccade to the target. When the system detects the eye making a saccade, the qualifying velocity criterion happens somewhere between two sample points, one on each side of the velocity criterion. This in turn gives us our temporal sampling error which will be, on average, half a sample in time (the expected mean of the uniform distribution [0,1]). As a result, our measured saccadic latency is the time from the stimulus onset to the true triggering of the saccade criterion *plus* the temporal sampling error. In Figure 2, this is the temporal error resulting from measuring the start of the trial, *a*, to the registered gaze criterion, *c*, and not the true saccadic latency that occurs between point *a* and point *b*. The temporal sampling error is the time between points *b* and *c*.

This form of temporal error is the same for a large group of measures we choose to call *one-point measures*. These measures have, by our definition, *either* the start criterion or the stop criterion of the measure defined by a frequency-independent system event and the other criterion defined by a sampled gaze criterion, but both criteria can *not* be determined by the same type of qualifying event.

The temporal sampling error caused by a finite sampling frequency where the true oculomotor events are uniformly distributed between sampling points can be described mathematically as follows. If the sampling interval spans the interval $[0, \frac{1}{f_s}]$, the sampling error can be described mathematically as

$$\varepsilon = t_{\text{Measured}} - t_{\text{True}}, \varepsilon \sim U(0, \frac{1}{f_s}) \quad (1)$$

where ε represents the error between the time (t_{True}) for the oculomotor movement and the time (t_{Measured}) when the movement was registered by an eye-tracker with sampling frequency f_s. $U(a,b)$ denotes the uniform distribution on the interval $[a,b]$

The net effect of the temporal sampling error on our desired measure depends on the particular measure. Specifically, the error will be positive or negative depending on whether the sampled gaze criterion constitutes the start criterion or the stop criterion in

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

the measure. This is because only sampled gaze criteria have temporal sampling errors, whereas system-generated criteria (which do not depend on a sampling frequency) have no such errors. The two possible outcomes that can result from a one-point measure can be expressed in the following way, where d is the duration estimated, s and g denotes system-generated and gaze sampled events respectively.

$$Overestimation: d+\varepsilon = (stop_g+\varepsilon) - start_s \quad (2a)$$

$$Underestimation: d-\varepsilon = stop_s - (start_g+\varepsilon) \quad (2b)$$

An overestimated d would result from any measure that use a sampled gaze property as a stop criterion, for example saccadic latency where the stop is the triggering of the saccade or time to target where the stop is the position of the eye inside the target area of interest. An underestimated d, on the other hand, would result from measures that uses a sampled gaze property as the start criterion, but a system generated event as the stop criterion. An example would be a measure we can call "time from decision", where we measure the time from when the participant gazed at a particular area of interest, to when the trial ends and the participant is forced to make a choice. In this case, the start event is sampled and the stop event is system generated (the end of the trial). This measure would result in an underestimated latency.

One-point measures that use a gaze sampled stop criterion are much more common than measures using only a gaze sampled start criterion. Therefore, we will primarily focus on the former type in this paper, but the equations and simulations can very easily be adjusted to accommodate those measures as well. Overestimated one-point measures have errors within the interval $[0, \frac{1}{f_s}]$ whereas the underestimated one-point measures are the mirror image with errors within $[-\frac{1}{f_s}, 0]$.

This means that, *on average*, the sampling error (whether it be overestimated or underestimated) will amount to half a sample worth of time. With a 50 Hz system, half a sample would amount to $\frac{1}{50}/2$ = 0.01 s = 10 ms.

The two-point temporal sampling error

There is another large group of measures that behave differently from the one-point measures. We choose to call them *two-point measures*, because they are qualified by *two* gaze-related criteria. These measures are both initiated *and* concluded by events determined from eye-movements and as such they contain sampling errors at both events. Their duration is determined by simply taking the end time of the event and subtracting the start time of the same event, and the resulting difference is the duration of the event. What differentiates these measures from one-point measures is that

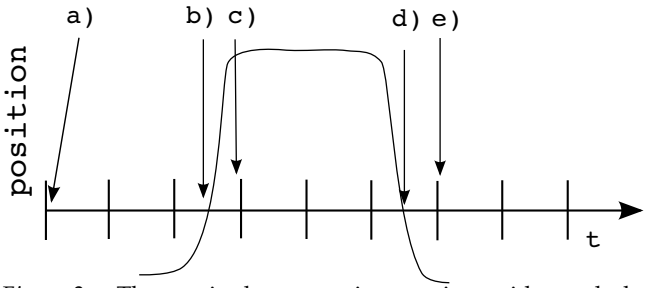

Figure 2. The x-axis shows continuous time with regularly occurring sampling points as brief vertical lines, starting from the start of the trial, *a*. The curved line crossing the x-axis shows the eye-movement fulfilling some gaze-based criterion, e.g. eye velocity or position in relation to an area of interest. This precise moment happens when the curve crosses the x-axis, *b*. This event is briefly after registered by the eye-tracker, in *c*. Similarly, another fulfilled criteria may happen later in *d*, but is registered only a while after, in *e*. The temporal sampling error is the difference between the true event and the registered event, e.g. *c-b* or *e-d*

two gaze-generated criteria results in two sampling errors within the same event. Consider Figure 2 and assume the gaze-based criteria to be the entering and the exit of an area of interest. Our measure will be time spent gazing at an area of interest in one visit, and we call this measure the dwell time. The dwell time can be expressed as $d^{dwell} = t^{dwell}_{stop} - t^{dwell}_{start}$, where d and t denote duration and point in time, respectively. In one end of this measure we have an entering event that occurs slightly before the registration of that event, and in the other end we have an exiting event which occurs slightly before the registration of the exit. The temporal sampling error can occur in both ends, i.e. at both qualifying criteria. The start of the measure is overestimated, yielding a later/higher start time, which results in a *shorter* dwell time (more is subtracted), but on the other hand we also overestimate the end of the measure, yielding a later/higher stop time and consequently a *longer* dwell time (more to subtract from). Thus, we can summarize the error of the dwell time, in this case, as $\varepsilon^{dwell} = \varepsilon_{stop} - \varepsilon_{start}$. As the events are equal in their distributions of the temporal sampling error, we expect that, on average, the net temporal sampling error will be zero - the two errors will cancel each other out. A net error of zero results when the error in estimating the start time and stop time of the two event criteria are exactly equal. However, the net error is not always zero, but is located between two extreme values $\varepsilon^{dwell} \in [-\frac{1}{f_s}, \frac{1}{f_s}]$. The first extreme value, a maximal underestimation of the dwell duration (-1 sample) occurs when we correctly capture the dwell stop as is occurs ($\varepsilon_{stop} = 0$), but we overestimate the dwell start by (almost) a complete sample ($\varepsilon_{start} = \frac{1}{f_s}$) – the real dwell starts immediately after we just sampled the eye, so we

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

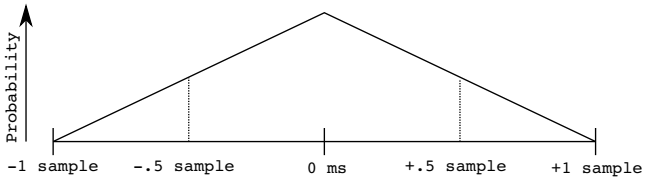

Figure 3. The triangular probability function for dwell time measurement error at a given sampling frequency, which is the distribution of the absolute difference between two uniform variables.

have to wait a complete sample to capture it inside the target area of interest. The net result ($\varepsilon^{dwell} = 0 - \frac{1}{f_s}$)) is an underestimation of the dwell duration by a complete sample worth of time. The other extreme point is the maximal overestimation of the dwell duration (+1 sample), which is the opposite event. Then, we correctly capture the dwell start as it occurs ($\varepsilon_{start} = 0$), but we overestimate the stop of the dwell because the triggering criterion happens immediately after we already sampled the eye, making us wait a complete sample for it ($\varepsilon_{stop} = \frac{1}{f_s}$). The net result ($\varepsilon^{dwell} = \frac{1}{f_s} - 0$) is an overestimation of the dwell duration by a complete sample of time.

A two-point sampling error is essentially the net result of subtracting a one-point temporal sampling error from another one-point error. For two-point sampling errors, the error distribution is no longer uniform, but assume the form in Figure 3. This distribution is the distribution of the absolute difference between two uniform variables.

Figure 3 shows us that at 50 Hz, the temporal error in a single dwell can be as large as 20 ms (a whole sample) in either direction, but that is very unlikely. In fact, a 0 ms error is much more likely, and the probability of the error being between -0.5 and +0.5 samples is as large as 75 %.

As the number (n) of dwells increases, the Central Limit Theorem (CLT) states that the temporal sampling error in the dwell time follows a Gaussian distribution with the average given by

$$\begin{aligned} \bar{\varepsilon}^{dwell} &= \frac{\varepsilon_1^{dwell} + \varepsilon_2^{dwell} + \ldots + \varepsilon_n^{dwell}}{n} \\ &\sim N(\mu, \frac{\sigma^2}{n}) \end{aligned} \tag{3}$$

where μ and $\sigma > 0$ are, respectively, the mean and standard deviation of the error ε^{dwell}.

Since the variance of a triangular distribution over an interval $[a,b]$ with mode c is $\frac{a^2+b^2+c^2-ab-ac-bc}{18}$, the distribution of the average error can be expressed as

$$\bar{\varepsilon}^{dwell} \sim N(0, \frac{1}{18nf_s^2}) \tag{4}$$

As we add data (e.g. dwells) the error distribution assumes an increasingly more pointy Gaussian distribution, making it less likely that the error will deviate from zero. Theoretically, with a large enough number of dwells, the average sampling error in your data goes to zero irrespective of sampling frequency.

From theory to practice

The theoretical relationship between sampling frequency and actual dependent variables in experiments may have large implications, especially for researchers using setups which favour naturalistic settings over higher sampling speeds, or researchers using video analyses. To test the theoretic predictions, we proceed to simulate these implications and to explore any real-world practical consequences this may have. The particular implications and questions to replicate and answer using these simulations are the following:

- How much data do we need in order to reduce the sampling error, for either a one-point or two-point measure, to a level comparable to a system sampling at 1000 Hz?
- How much more data do we need in order to overcome the sampling error and get the average t-test significant, for either a one-point or two-point measure?
- Is the empirical distribution of a two-point temporal sampling error the same as the distribution predicted in Figure 3?
- How large are the effects of temporal sampling errors in relation to the effects of sampling frequency on event detection algorithms and their particular parameter settings? Which error is worse?

Simulation 1 - one-point error reduction

The aim of this simulation is to explore how much data we need to reduce the variance of the temporal sampling error for one-point measures to a level where it is easy to estimate and compensate for. The goal is that 95 % of all temporal sampling errors should be within the same 1 ms span. The 1 ms span is selected as a baseline because eye-tracking research seldom tries to estimate effects under 1 ms. A 1 ms error would correspond to the *maximal* one-point temporal sampling error caused by a 1000 Hz eye-tracker. This simulation tells us at what data amount we can simply subtract the expected temporal sampling error from the means to get the true estimate.

Procedure

The acceptance level is an average deviation from the expected mean of the temporal sampling error within 1 ms for 95 % of all iterations. For example, at 50 Hz, the expected mean of the sampling error is $\frac{1}{50}/2 = 0.01$ s = 10 ms. If the mean of the tested data

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

amount is within the interval [9.5, 10.5] for 95 % of the iterations and no smaller data amount fulfill the same criteria, we accept that data amount as the optimal level for reducing the one-point temporal sampling error.

The pseudo-code for Simulation 1 describes how the simulation was performed at implementation level.

Simulation 1 Pseudo code for Simulation 1.

```
for f_s = 10 to 2000 in steps of 10 (sampling frequency)
do
  i = 1 (data amount, i.e. number of one-point measures)
  E(ε) = 1/f_s /2 (expected temporal sampling error)
  while 1 do
    Generate 10000 one-point error vectors ε⃗ = {ε_1, ε_2, ..., ε_i} from eq. (1) of length i
    if 95 % of the ε⃗ have means within [E(ε) − 0.5, E(ε) + 0.5)] ms then
      store i and f_s
      Break while-loop
    else
      i = i + 1
    end if
  end while
end for
```

Results & Discussion

Figure 4 shows the relationship between sampling frequency and the required data amounts to contain the temporal error within 1 ms of the expected mean of the error. The data requirements to reduce the temporal errors are very steep for lower sampling frequencies, but as the frequencies approach 200 Hz, the requirements level out and differ very little between 200 Hz and 1000 Hz.

We managed to fit the simulation results near perfectly (r^2 = .99) to Equation (5) where N is the data points required, c is the constant 1208500 and f_s is the sampling frequency.

$$N = c f_s^{-2} \qquad (5)$$
$$f_s = \sqrt[2]{\frac{c}{N}}$$

Given the sampling frequency, we can solve for the minimum number of data points required to contain the temporal sampling error within 1 ms of the expected mean of the error. Similarly, if we have the number of data points, we can solve for the minimum sampling frequency needed in order to contain the temporal sampling error within 1 ms of its expected mean.

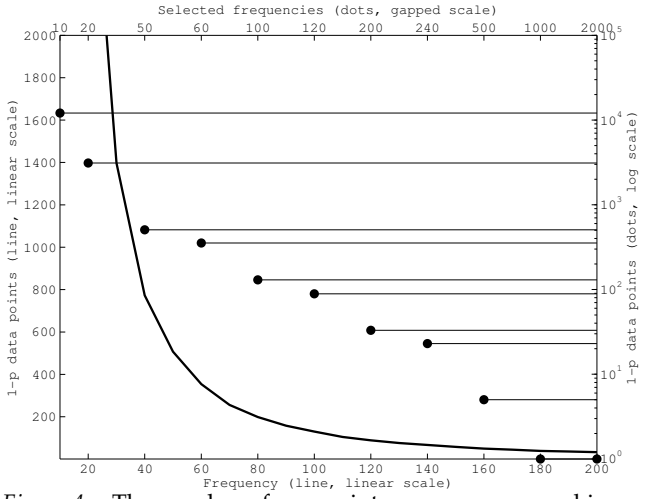

Figure 4. The number of one-point measures we need in order for the mean temporal error to be less than 1 ms. The line shows the fixations needed for all simulated frequencies (left and bottom scales). The dots show the fixations needed for typical frequencies of modern eye-trackers (top and right scales, base 10 log-transformed).

Simulation 2 - one-point data compensation

However, we are often not interested in reducing the temporal sampling error to within 1 ms, but just getting a significant difference between an experimental condition and a control condition. Therefore, a more practical question would be: Given a particular effect magnitude, how much data (data points of a one-point measure) do we need in order to find a significant difference between the two conditions?

Procedure

This simulation tested every sampling frequency from 10 Hz to 2000 Hz in steps of 10 Hz and fixed (as in no variance) effect sizes of 5, 20 and 50 ms. Random one-point sampling errors were generated into two sets: one containing only the base error, and the other containing the base error with the added constant effect. Data set sizes were gradually increased one data point at a time and data were generated until the two sets were significantly different at a 95 % confidence level using a two-sample t-test. Each unique parameter combination was repeated 10,000 times and t-test results were only accepted if 95 % of all 10,000 t-tests were significant to avoid a multiple comparisons problem.

Results & Discussion

The results, which are shown in Figure 5, show that on a log-log scale, there seems to be a perfectly linear relationship between the sampling frequency and

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

Simulation 2 Pseudo code for Simulation 2. c is the effect size.

```
for f_s = 10 to 2000 in steps of 10 do
  for c = 5, 10, 20 ms do
    i = 1
    while 1 do
      Generate 10000 error vectors pairs ε⃗_1 = {ε_{1,1}, ε_{1,2}, ..., ε_{1,i}} and ε⃗_2 = {ε_{2,1} + c, ε_{2,2} + c, ..., ε_{2,i} + c} from eq. (1) + c of length i.
      if 95% of the pairs are significantly different at a 5 % level then
        store i, f_s and c
        Break while-loop
      else
        i = i + 1
      end if
    end while
  end for
end for
```

the number of data points you need. As you investigate smaller effects, the curve will shift outwards and increase your data requirements. For small effects at 5 ms, a 100 Hz system needs around 10 data points. Lowering the speed to 50 Hz, increases the data requirements to around 40 data points. It is important to point out that these effects are constant, i.e. always 5/20/50 ms for each measure in one of the vectors. Real effects are seldom constant, but rather normal in their distribution, which means these results represent an optimistic minimal-requirements case for these effect magnitudes. The absolute levels are not the main focus here, but rather the relation between sampling frequency and data requirements.

Simulation 3 - two-point temporal error reduction

In this simulation, we investigate how many data points from a two-point measure we need, given a particular sampling frequency, in order for the temporal sampling error be limited to maximally 1 ms (we select the same span as in the one-point simulation - for comparison). This is similar to convolving the distribution in Equation (4) until it reaches such a narrow Gaussian form that it is very unlikely (p <.05) that the sampling error is within 1 ms (.5 ms in either direction of the mean).

Procedure

The same procedure as in Simulation 1 was used, but the two-point temporal sampling error was instead calculated by Equation (3). Note that the expected temporal sampling error for a two-point measures is zero, so the resulting error distribution will be centered on zero.

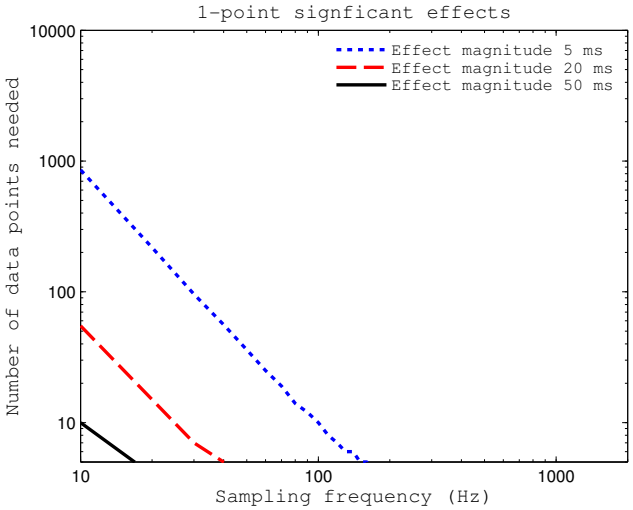

Figure 5. The number of one-point measures we need in order get a two-sample t-test significant at the 95 % level for three different effect magnitudes. Values below 5 data points are not shown as t-tests are not reliable for such small samples.

Results & Discussion

The results in Figure 6 indicate that for very low sampling frequencies, the data requirements to greatly reduce the sampling errors are enormous, but these requirements drop off very quickly as the frequency increases. At around 200 Hz or above, there is little marginal benefit of higher sampling frequencies with regard to reducing sampling errors. Furthermore, we managed to fit the data near perfectly (r^2 = 1.00) using Equation (5), where N is the data points required, f_s is the sampling frequency, but with c as the constant 2429400 (which differs from the constant for solving one-point errors).

Given the sampling frequency, we can solve for the minimum number of data points required to contain the temporal sampling error within 1 ms of the expected mean of the error. Similarly, if we have the number of data points, we can solve for the minimum sampling frequency needed in order to contain the temporal sampling error within 1 ms of its expected mean.

Simulation 4 - two-point data compensation

Of course, the typical researcher is often not interested in completely cancelling this sampling error, but rather to show that her experimental manipulation has a statistically significant effect. In this simulation we investigate, given a particular sampling frequency and a particular effect magnitude, how many pairs of data points from a two-point measure will suffice in order to achieve a significant two-sample t-test on the average comparison. This simulation will show how large a

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

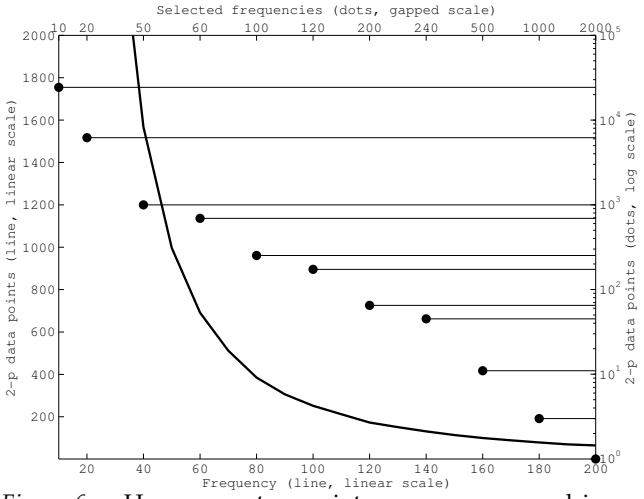

Figure 6. How many two-point measures we need in order for the mean temporal noise to be less than 1 ms. The line shows the fixations needed for all simulated frequencies (left and bottom scales). The dots show the fixations needed for typical frequencies of modern eye-trackers (top and right scales, base 10 log-transformed).

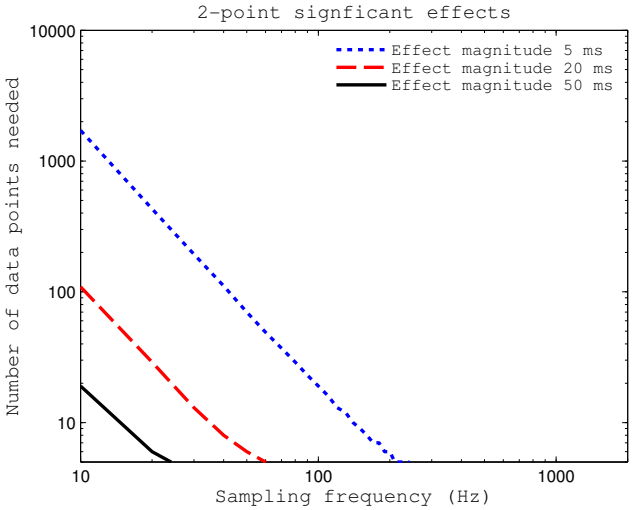

Figure 7. Number of two-point measures needed for two samples to be significantly different at various effect magnitudes. The two samples compared are one with a base two-point sampling error and another with a base sampling error plus the added constant effect. Data below 5 data points are not shown as t-tests are not reliable for very small data sets.

role this sampling error has in adding noise to two data sets and how this error consequently affects hypothesis tests.

Procedure

The same procedure as in Simulation 2 was used, but the errors were generated for a two-point measure using Equation (4).

Results & Discussion

Results in Figure 7 show the same relation between sampling frequency and data requirements as Simulation 2, which was the same simulation for one-point measures. However, the absolute values are slightly different, reflecting the fact that two-point temporal sampling errors span a larger interval, and consequently require more data to obtain a mean near zero. The same reservation as for the one-point measures remain – that this reflects an optimistic minimal-requirements case only.

Simulation 5 - real-world data

In this simulation, we resample real eye-tracking data from a reading task to quantify the sampling error and verify the predicted shape of the two-point error distribution in Figure 3. This is done in order to show that this is not only a purely theoretical effect, but it can also affect actual recorded data

Procedure

We used real eye-tracking data from the large reading experiment described in Nyström and Holmqvist (2010). In short, eye-movements were recorded at 1250 Hz while University students read texts on a computer screen. The text was divided into 16 screens (images), and for each screen we defined an area of interest around a single high-frequency word near the center of the screen. We used this area of interest to calculate the time spent gazing inside this area in one single visit from entry to exit, referred to as a dwell, and the resulting dwell time. We estimated dwell time using the original 1250 Hz raw data and only included dwells that were longer than 50 ms (similar to the shortest fixations, see e.g. Rayner, 1998:376) and were separated by at least 20 ms. The 20 ms criterion corresponds to the minimum duration of a saccade, which is 10 ms according to Nyström and Holmqvist (2010), rounded up to 20 ms to equal a full 50 Hz sample. This allowed us to ignore high-frequency noise such as visits due to passing saccades and eye-tracker imprecision. We then downsampled this data to 50 Hz data by using every 25^{th} coordinate pair. The 1250 Hz data function as a baseline that we, for sake of argument, assume are identical to the objective sampling of an unlimited sampling frequency. The differences that arise are thus due to the longer sampling intervals of the 50 Hz system, essentially showing the sampling error of a 50 Hz system relative to a 1250 Hz system.

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

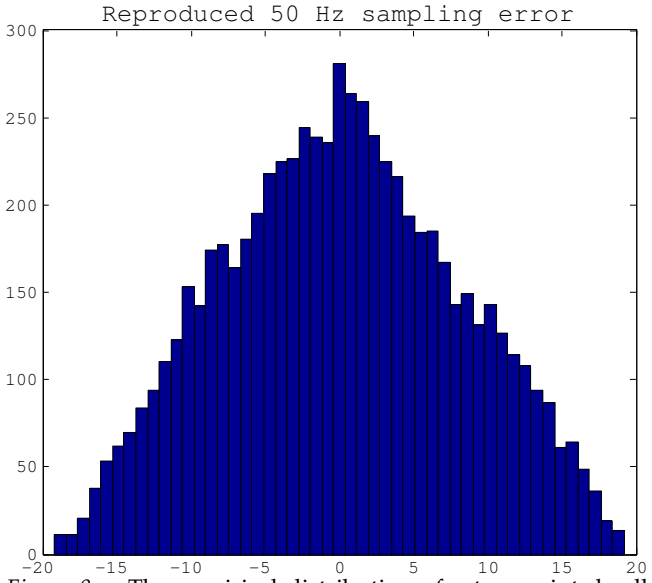

Figure 8. The empirical distribution of a two-point dwell time sampling error.

Results & Discussion

Figure 8 shows our obtained sampling error of a two-point measure, which in our case was dwell time. This replicates the distribution predicted in Figure 3. The isolated two-point temporal sampling errors are constrained within $\pm$ 1 sample, and show that measures generated from real data are subject to the effects demonstrated in Simulation 2. The algorithm identified 6804 dwells, and they were on average 224.94 ms (std 141.60) for the 1250 Hz system, and 224.84 ms (std 141.88) for the 50 Hz system.

Simulation 6 - event detection

Up to this point, we have assumed that the only source of error when estimating dependent measures is the temporal sampling error resulting from the limited sampling frequency of the eye-tracker. In practice, however, fixations and saccades are identified by event detection algorithms, and sampling frequency is only one of the factors that affect how reliably fixations and/or saccades are detected. Other factors include the precision of the data, methods used to calculate eye-movement velocity and acceleration, type of algorithm, algorithmic settings, and thresholds (Salvucci & Goldberg, 2000; Shic, Scassellati, & Chawarska, 2008; Blignaut, 2009; Nyström & Holmqvist, 2010). The aim of this simulation is not to thoroughly examine all the effects that sampling frequency may have on event detection, which is a large question itself, but rather to put the sampling error in perspective against another error source which also depend on sampling frequency.

Procedure

To investigate the effect sampling frequency has on event detection, data collected at 1250 Hz (same as in Simulation 3) was used as a baseline, and then compared to the same data resampled to 250 and 50 Hz, respectively. Resampling was made by first downsampling the baseline data with factors 5 and 25, and then upsampling it back to its original size using nearest-neighbour interpolation. This way, three data sets of the same size were generated.

Traditional methods for fixation and saccade detection typically use fixed thresholds to detect saccades and/or fixations (e.g., everything above/below a certain velocity threshold is a saccade/fixation) . Since the resampling process changes the signal characteristics, these thresholds would need to be modified in order to accurately detect events. This is particularly true for the resampled 50 Hz data. Based upon these arguments, fixation and saccade detection was performed with the algorithm by Nyström and Holmqvist (2010) using a peak detection threshold of $\dot{\theta}_{PT} = \mu + 4\sigma$, which objectively adapts with the signal characteristics. All other parameters were fixed.

Results & Discussion

As shown in Table 1, a sampling frequency of 250 Hz is almost identical to one of 1250 Hz in terms of both fixation and saccade duration. At 50 Hz, however, the measures start to diverge, and saccade durations are more sensitive to sampling frequency than fixation durations are. In fact, even with more than 25,000 saccades, the durations from the 50 Hz system are not identical to the durations from the 1250 Hz or 250 Hz systems, indicating that whatever temporal error introduced by the event detection process the error is not centered on zero. Checking the two-point data requirements in Figure 6, a 50 Hz system should require about 1000 data point to contain the error within 1 ms of its expected mean for 95% of all cases. In this simulation, we have about 25,000 saccades, but the difference compared to the 250 Hz and 1000 Hz system is 5 ms. This means we are dealing with two types of temporal sampling errors. If only a sampling error of the type explained and tested in the rest of this article had been present in the fixation/saccade detection, then at almost 20,000 fixations and 25,000 saccades, the error should be virtually zero. As this is not the case, there must be other errors causes by the sampling frequency at work. A lesson from this is that it may pay off better to focus on selecting a reliable event detection algorithm rather than worrying about temporal sampling errors, especially if the amount of data available is not a problem.

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

Sampling frequency (Hz)	1250	250	50
Fixations (number)	19656	19793	19955
Fixation duration, M±SD, (ms)	177±83	177±83	175±85
Saccades (number)	24684	24768	24835
Saccade duration, M±SD, (ms)	43±19	43±19	48±22

Table 1
Effect of sampling frequency on number and duration of saccades and fixations (reading data from 10 participants).

General discussion

We calculated and simulated the relationship between sampling frequency and data requirements. Quadratic curves were fitted to the data, which can be summarized in the following rule of thumb: a doubling in sampling frequency allows for only a fourth of the data, in order to maintain a constant sampling error. Of course, this can also be rephrased as: if you halve the sampling frequency, you will need four times as much data to maintain the same temporal sampling error. However, these advice concern only the temporal sampling error, and you still need data to reach adequate power to confidently identify your experimental effects of interest.

Examining the data which was fit to Equation (5), we also find that the constant, c, is around twice as large for two-point measures compared to one-point measures. The implication is that, as a rule of thumb, we need approximately twice the amount of two-point measures in order to contain the temporal sampling error within a span that is equal in size that of one-point measures.

Two-point measures are a particular group of measures, such as fixation duration, dwell time and saccade duration. These measures have both an initiating and a concluding event generated by gaze behaviours, and as such have sampling errors at two separate points. These errors even out on average, and more importantly, more data makes it more likely that the average temporal sampling error will be close to zero.

One-point measures, such as saccadic latency, antisaccadic correction latencies and time to target, however, lack this property. The error of those measures does not reduce, but at least it becomes more stable and converge at an error of half a sample in time. One-point measures are all measures that involve a single qualifying gaze behaviour, and the other qualifying events should be system generated (independent of sampling frequency). They are therefore very often latency measures. It is possible to correct this mis-estimation of one-point measures by simply subtracting or adding half a sample worth of time to center the temporal sampling error on zero. Recording more data will *then* mean that the one-point error will be close to zero. Depending on whether the start or the stop criterion of the measure is gaze sampled, the correction procedure will be adding or subtracting, respectively.

For some labs, sampling errors are not a problem because all their research are performed using 1000+ Hz eye-trackers and these sampling errors pose no practical problems. However, for natural studies where the participants should be able to move their heads, perhaps reading a newspaper, working with several monitors or even naturally walking around outside the lab, this may be an issue. For naturalistic eye-tracking studies, current eye-tracking systems record at most at 250 Hz, but often much slower such as 50 Hz, and manual video analysis at 25 fps/Hz is still common. Similarly, some researchers focus exclusively on low-cost, and hence slower, cameras in order to provide accessible systems for more users. We believe it is important to understand how large these sampling-generated errors are and at what point they cease to be a problem. Also, some research questions allow only limited amounts of data to be recorded, making it important to select a system that will minimize the sampling errors.

Fortunately, sampling error is not a practical problem given eye-trackers that operate at roughly 200 Hz or above (depending on available data amounts). Often, event detection algorithms pose more of a problem than sampling errors. This is mainly because event detection errors are not centered on zero, where more data negates the errors, and because the event detection errors are not completely predictable, neither in direction nor magnitude (and depending on type of detection algorithm). Our final simulation attempted to put the variation between event measures in perspective against the magnitudes of sampling errors. For truly accurate estimation of measures, it can be beneficial to avoid an event detection process altogether, unless it is explicitly required by a measure focused on a particular oculomotor event such as a fixation or a saccade.

It should be noted that the temporal sampling error described in this paper has implications for all users of eye-trackers, though some are more affected than others. Two-point temporal measures, such as fixation durations and duration of visits, are possible to use even with a low-speed eye-tracker. However, it depends entirely on the researcher being able to add more data. This translates into more participants and more trials, which are not equally easy to add for all types of experiments. For example, a supermarket study analysed by video (equal to 25 Hz) involves more effort to recruit and test participants, and unrestricted participants produce the number of visits on products they do, no more, no less. To produce more data, this may involve only recruiting shoppers with long grocery lists, and avoid the single item shoppers, or generally just adding more participants. For manually analysed data, adding more data is an alternative preferably avoided.

Similarly, researchers investigating low-cost cameras for gaze interfacing are very restricted in the ability to acquire more data. If they are interested in using gaze for, e.g., dwell-time based triggers on the screen, then

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

they only have a single data point, one dwell, to work with. In this case, it is unclear whether there exists some way to get around the problem at all.

However, even though all researchers may not be able to reduce their temporal sampling error in the way described in this paper, at least all should be able to calculate this error and decide for themselves whether it is a problem or not. It may be fruitful for a gaze-interfacing researcher to further investigate how participants handle this temporal sampling error. Do gaze-interface users experience an annoying element of uncertainty in using dwell-time based triggers due to the temporal sampling error, and in that case, at what sampling frequency does that uncertainty cease to be annoying?

The findings of this study has the greatest impact for researchers who have traded system speed for ecological validity in their setups. For example, using a head-mounted eye-tracker or a remote eye-tracker, which typically (but not necessarily) have speeds below 200 Hz. This is especially important for experiments using special populations, e.g. clinical groups, children or even primates, which may put a limit on the number of trials that can be recorded.

Our aim has been to break down sampling-related errors in an accessible form for all users of eye-trackers, and we suggest the following heuristics for researchers and reviewers:

- Are you using one-point temporal measures such as saccadic latency or time to target? On average, the measures are mis-estimated by half a sample of time. This error decreases linearly as sampling frequency increases. For example, using a 50 Hz system will on average mis-estimate the duration with 10 ms, and a 100 Hz system will on average mis-estimate with 5 ms. Is the temporal sampling error a problem? Either check Figures 4 and 5 , or use Equation (5) to calculate the adequate (equivalent to a 1000 Hz system) number of data points needed or adequate sampling frequency. If you want true estimates of a one-point error, you need to subtract or add half a sample of time from them to center them on zero (depending on the partcular measure). Use Equation (5) to calculate the requirements for you to estimate them within 1 ms.
- Is the two-point temporal sampling error really a problem your experiment? Either check sampling frequency and data amount against the graph in Figure 6, or calculate the data requirements using Equation (5). For example, if you have a sampling frequency of 60, you need at least 675 data points. On the other hand, if you are a reviewer and read an article that has used a given number of data points, you can use Equation (5) (use the correct constant!) to calculate the minimum sampling frequency they should have used. For example, 24 participants tested on 30 trials each, where every trial produce one (1) data point (e.g., one dwell duration), then the total number of data points are $24 \cdot 30 \cdot 1 = 720$. This translates into a minimum sampling frequency of 58 Hz. At 58 Hz, the temporal sampling error is not eliminated, but equivalent to that of a 1000 Hz system and in effect negligible.
- Are you considering using a slower, but more naturalistic, system set-up? Use the quadratic relationship to calculate the increased data requirements. Halving the frequency means you have to increase the data requirements a fourfold to maintain the same temporal sampling error.
- Are you planning to buy an eye-tracker? Check Figures 4 and 6 to see just how much more data you need to record given the sampling frequency of your different candidate systems. This may be an issue if the sampling frequency is low and/or you are limited in the number of data points (trials) you can record per participant, e.g. if you use babies or perhaps primates.
- Do you want to compare studies two studies using different sampling frequencies? You can then check the amount of data used against Equation (5) or Figures 4 & 6, and see if the temporal sampling error is contained within 1 ms. If this is the case for both studies, then it is safe to compare estimates. If any of the studies have used one-point measures, such as saccadic latency, then you may have to subtract or add half a sample of time to correct the estimates.

Conclusion

We found that one-point measures, such as various latencies, have a temporal error distribution centered on half a sample in time, but can be centered on zero by deducting or adding the corresponding amount of time. Two-point measures, such as fixation and saccade durations, have sampling errors centered on zero. Both measure groups will be more accurately estimated by adding more data (assuming one-point measures have been centered on zero). Finally, there is a quadratic relationship between sampling frequency and data amount, where a doubling of sampling frequency lowers data requirements to one fourth if the goal is to maintain the same average temporal sampling error. Another rule of thumb is that two-point measures, such as fixation durations, require around twice the amount of data compared to one-point measures to contain the temporal sampling error within the same span.

References

Blignaut, P. (2009). Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception & Psychophysics*, *71*, 881-895.

Enright, J. T. (1998). Estimating peak velocity of rapid eye movements from video recordings. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 349-353.

Green, P. (2002). Human factors in traffic safety. In R. E. Dewar & P. L. Olson (Eds.), (p. 77-110). Tucson, AZ: Lawyers & Judges Publishing Company, Inc.

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692

Inchingolo, P., & Spanio, M. (1985). On the identification and analysis of saccadic eye movements-a quantitative study of the processing procedures. *IEEE Transactions on Biomedial Engineering*, *32*(9), 683-695.

Juhola, M., Jäntti, V., & Pyykkö, I. (1985). Effect of sampling frequencies on computation of the maximum velocity of saccadic eye movements. *Biological Cybernetics*, *53*(2), 67–72.

Kagan, I., Gur, M., & Snodderly, D. M. (2008, 11). Saccades and drifts differentially modulate neuronal activity in V1: Effects of retinal image motion, position, and extraretinal influences. *J. Vis.*, *8*(14), 1-25. Available from http://journalofvision.org/8/14/19/

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, *17*(6), 578–586.

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*, 188-204.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Salvucci, D., & Goldberg, J. (2000). Identifying fixations and saccades in eyetracking protocols. In *Etra '00: Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 71–78). New York, NY, USA: ACM Press.

Shannon, C. (1949). Communication in the presence of noise. *Proceedings of the IRE*, *37*(1), 10–21.

Shic, F., Scassellati, B., & Chawarska, K. (2008). The incomplete fixation measure. In *Etra '08: Proceedings of the 2008 symposium on eye tracking research & applications* (pp. 111–114). New York, NY, USA: ACM Press. Available from http://dx.doi.org/10.1145/1344471.1344500

DOI 10.16910/jemr.3.3.6

ISSN 1995-8692