Behavior Research Methods
https://doi.org/10.3758/s13428-021-01782-4

Factors affecting inter-rater agreement in human classification of eye movements: a comparison of three datasets

Lee Friedman[1] · Vladyslav Prokopenko[1] · Shagen Djanian[1,2] · Dmytro Katrychuk[1] · Oleg V. Komogortsev[1]

Accepted: 22 December 2021

Abstract

Manual classification of eye-movements is used in research and as a basis for comparison with automatic algorithms in the development phase. However, human classification will not be useful if it is unreliable and unrepeatable. Therefore, it is important to know what factors might influence and enhance the accuracy and reliability of human classification of eye-movements. In this report we compare three datasets of human manual classification, two from earlier datasets and one, our own dataset, which we present here for the first time. For inter-rater reliability, we assess both the event-level F1-score and sample-level Cohen's κ, across groups of raters. The report points to several possible influences on human classification reliability: eye-tracker quality, use of head restraint, characteristics of the recorded subjects, the availability of detailed scoring rules, and the characteristics and training of the raters.

Keywords Eye-movements · Manual classification · Sample-level agreement · Event-level agreement · Cohen's Kappa · F1-score

Introduction

Manual classification is used to detect different eye movement types and is also used as a comparison to automatic methods (Agtzidis et al. (2020), Andersson et al. (2017), Dar et al. (2021a), Fuhl and Kasneci (2021), Hooge et al. (2018), Jongerius et al. (2021), Korda et al. (2015), Kothari et al. (2020), Larsson et al. (2013), Startsev et al. (2019b), Stuart et al. (2018),Vargas-Cuentas et al. (2017), Venker et al. (2020), Wadehn et al. (2018a), Zemblys et al. (2018), Zemblys et al. (2019)). In addition, manual coding is used to train machine-learning models (Zemblys et al. (2018, 2019)). Manual classification will not be very useful, however, if there is little inter-rater classification agreement. In this study, we look at some qualities and characteristics of studies that can affect inter-rater agreement. Our hope is that this analysis will help researchers develop more meaningful and repeatable manual classification approaches.

To conduct this study, we needed to find manually annotated datasets that contained periods of fixation, saccades and possibly post-saccadic oscillations and did not contain a substantial portion of time in smooth pursuit. We were also only interested in datasets where the data was publicly available. We performed a Medline search with the following syntax:

> eye movement AND manual AND (classification OR annotated OR labelled OR coded) NOT (sleep OR NREM OR REM).

This search produced 97 articles, and based on the title and abstract we thought that 12 might be candidates. One we knew of by other means (Fuhl & Kasneci, 2021). The studies are listed in Table 1 along with an included/excluded decision and reason. We included two studies from this list in the present analysis (Hooge et al., 2018; Larsson et al., 2013)[1]. These 2 studies along with the study we report herein comprise the 3 datasets employed in the present analysis.

The Hooge et al. (2018) dataset was conducted to determine if human classification can be a "gold standard".

✉ Lee Friedman
lfriedman10@gmail.com

[1] Derrick M5, Department of Computer Science, Texas State University, 601 University Drive, San Marcos, Texas, 78640, USA

[2] Present address: Department of Computer Science, Aalborg University, Selma Lagerlofs Vej 300, 9220 Aalborg East, Denmark

[1] A correction to Hooge et al. (2018) was published in Hooge et al. (2021)

Published online: 11 April 2022

Table 1 List Of Potential Studies

Study	Include/Exclude
Jongerius et al. (2021)	Classified areas of interest, not eye movement types.
Fuhl and Kasneci (2021)	Complex stimuli with lots of smooth pursuit mixed with fixation and saccades.
Dar et al. (2021b)	Excluded: Used data from Larsson et al. (2013)
Kothari et al. (2020)	Excluded:Head-free gaze behavior is significantly more complex with a wide range of behaviors to be classified.
Agtzidis et al. (2020)	Excluded: Complex video stimuli with much smooth pursuit mixed with fixation and saccades. One human rater.
Venker et al. (2020)	Excluded: Manual coding of video, not eye movement signals.
Startsev et al. (2019b)	Excluded: Complex video stimuli with much smooth pursuit mixed with fixation and saccades.
Stuart et al. (2018)	Excluded: Manual classification of images, not eye-movement signals.
Wadehn et al. (2018b)	Excluded: Used data from Larsson et al. (2013).
Hooge et al. (2018)	Included
Andersson et al. (2017)	Excluded: Used data from Larsson et al. (2013)
Vargas-Cuentas et al. (2017)	Excluded: Manually classified images and videos, not eye-movement signals.
Korda et al. (2015)	Excluded: Data not available.
Larsson et al. (2013)	Included

It was based on comparatively low quality data: a relatively low quality eye-tracker (see Table 2)[2], no head restraint employed, 87% of recordings were from infant subjects (57% of total recording time was from infants). Also Hooge et al. (2018) did not rely on detailed scoring rules and used no rater training but did use experts. Only fixation classification was performed. The Larsson et al. (2013) dataset was not originally intended to be used to evaluate inter-rater reliability but was used as a basis for comparison to a new automatic classification system. Nonetheless, it can be used for the former purpose. As an evaluation of human classification performance, it was based on data from a high-end eye-tracker and it employed only adult subjects. It was classified by two raters who were highly regarded experts in eye movement classification (Marcus Nyström and Richard Andersson). The present dataset used data from a high-end eye-tracker, with head restraint. It was based on young healthy adult subjects, employs detailed scoring rules[3] and very extensive rater training. The latter 2 datasets classified fixations, saccades and PSOs.

We will compare these datasets in terms of inter-rater reliability of manual classification using sample-level Cohen's κ, as well as event level F1-score analyses. Cohen's κ is frequently used in such studies (Dar et al. (2021a), Hooge et al. (2018), Kothari et al. (2020), Larsson et al. (2013), Startsev et al. (2019b), Zemblys et al. (2018), Zemblys et al. (2019)) for sample-level agreement[4]. Recent studies have employed the F1-score to assess event-level agreement (Agtzidis et al. (2020), Hooge et al. (2018), Kothari et al. (2020) Startsev et al. (2019c)).

There is a tutorial[5] on the F1-score that we find useful and recommend entitled "A Look at Precision, Recall, and F1-Score". According to the creator (Teemu Kanstrén):

> The F1-Score is a measure combining both precision and recall. It is generally described as the harmonic mean of the two. Harmonic mean is just another way to calculate an "average" of values, generally described as more suitable for ratios (such as precision and recall) than the traditional arithmetic mean. The formula used for F1-score in this case is:
>
> $2 * (Precision * Recall)/(Precision + Recall)$
>
> The idea is to provide a single metric that weights the two ratios (precision and recall) in a balanced way, requiring both to have a higher value for the F1-score value to rise.[6]

Methods

The (Hooge et al., 2018) dataset

We will use the words of Hooge et al. (2018) to describe their dataset.

[2]Note that we are not saying that the Tobii TX 300 is low quality compared to the universe of eye-trackers. Only that, of the 3 eye-trackers involved in the current report, it is the lowest quality.

[3]Available at https://digital.library.txstate.edu/handle/10877/13373

[4]Percent agreement, which does not take into account the role of chance, is deprecated since the introduction of Cohen's κ.

[5]https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

[6]For a presentation of the the problems that can arise when applying a kappa versus the F1-score to event related data, see Friedman (2020).

Table 2 Comparing Eye Trackers on Data Quality

Eye-Tracker	Sampling Rate	Precision Ranking Wang et al. (2017)	Accuracy Ranking Holmqvist (2017)	Precision Ranking Holmqvist (2017)
EyeLink 1000	1000 Hz	1st or 2nd of 10	1st of 12	2nd of 13
SMI HighSpeed 1250	500 Hz	1st or 2nd of 10	3rd of 12	3rd of 13
Tobii TX 300	300 Hz	8th of 10	4th of 12	5th of 13

> Twelve experienced but untrained human coders classified fixations in 6 min of adult and infant eye-tracking data. (page 1864)
>
> The eye-tracking stimulus set consists of 70 trials of eye-tracking data measured with a Tobii TX300 at 300 Hz. We used eye-tracking data measured from the left eye. Ten of the 70 trials contained 150.1 s of eye-tracking data of two adults looking at Roy Hessels's holiday pictures taken in the arctic area around Tromsø, Norway. The other 60 trials contained 202.1 s of eye-tracking data of infants performing a search task Hooge et al. (2018). (page 1867)

Note that Hooge et al. (2018) employed 60 recordings from 60 individual infants and employed 10 recordings from 2 adults. Recordings of the same person are not independent, whereas recordings from different subjects are independent. We consider that mixing of independent and dependent observations in one analysis confounds the interpretation of the results.

With 150.1 s and 10 adults, there were, on average 15.01 seconds of recording for each adult. For 202.1 seconds and 60 infants, there were, on average, 3.6 seconds per infant recording.

Compared to the other eye-trackers in this study, the Tobii TX 300 was relatively low-ranked (see Table 2).

It is noteworthy that for this dataset, the actual number of adults whose recordings were evaluated was 2. For the Lund dataset, described below, we were able to analyze the recordings from 12 different subjects. For the present dataset, described below, we analyzed recordings from 19 distinct subjects. So, in the Hooge et al. (2018) dataset there were 2 adult subjects who provided a total of 10 separate recordings, whereas in the other datasets every recording analyzed was from a unique subject. This should contribute to the greater generality of the results from latter 2 datasets as opposed to the Hooge et al. (2018) dataset.

> Trials of both the adult and the infant eye-tracking datasets were presented in random order on a 24-in. TFT screen (1,920 *X* 1,200 pixels). The vertical axis of the position signals was fixed (respectively, 0–1,920 and 0–1,080 pixels, since measurements were done on the HD screen of the TX300). ...Each screen showed 1 s of data and contained the last 250 ms of the previous display (to provide context) and 750 ms new data at a time. (page 1867)

The infant subjects presented particular challenges. For example, 11 of the 60 infant recordings had more than 50% missing data, and one infant had 94% missing data. Also, in their Table 2, they indicate that for every RMS noise estimate, RMS error was higher in infants than adults[7]. For this reason, we chose to analyze only the 10 adult recordings from that dataset. For examples of the adult recordings in Hooge et al. (2018), see Friedman (2020). The effect of this choice is likely to be that the inter-rater reliability we report for this study is higher than that which would have obtained if we had included the infant recordings.

In a personal communication with Dr. Hooge,we were informed that no forehead or chinrest was used to collect this data. For a description of some of the data quality issues that arise from a lack of head restraint see Niehorster et al. (2018).

> "We engaged 13 eye-tracking researchers in the fixation labeling task. We removed one human coder from the analysis because we found out he had never looked at raw data before. The remaining 12 coders are members from different research groups; details about them may be found in Table 1." (page 1867)

The coders were considered experts. We will refer to these "coders" as "raters" henceforth.[8] We did employ the recently corrected data made available in Hooge et al. (2021).

[7] Particularly noteworthy is the maximum RMS, which was a factor of 3 larger in infants compared to adults.

[8] For additional data-quality issues with this dataset, see Friedman (2020)

The Lund dataset

The Lund Dataset was first described in Larsson et al. (2013). The eye-tracking signals were collected from 31 participants, including students and laboratory personnel. The mean age of the participants was 31.2 +/- 9.9 (M +/- SD) years. The signals were recorded binocularly (but only the right eye data was analyzed) with the iView X HiSpeed 1250 eye-tracker from SensoMotoric Instruments (Berlin, Germany). Different types of stimuli were presented. For present purposes, we used only the data where subjects were viewing static images for 10 seconds. We only used datasets collected at 500Hz[9]. The recordings were classified into fixations, saccades, post-saccadic oscillations, smooth pursuit, blinks and undefined by two raters, (MN = Marcus Nyström, RA = Richard Andersson). These two raters are considered to be international experts in the classification of eye movements. We had 12 pairs of recordings, with one of each pair classified by rater MN and the other classified by rater RA. Although these raters did classify smooth pursuit, in general, smooth pursuit was very rare (0.92% of total recording time for rater MN,4.78% for RA) when subjects were viewing static images, as would be expected.

The two coders labeled the data samples manually and separately, using the same Matlab GUI. This GUI contained several panels that together showed, for a stretch of data in time, the current x- and y- positions (in pixels), the velocity (point-to-point, in degrees per second), and a scatterplot representations of the data. Additionally, the GUI also showed a zoomed in portion of the sample positions, as well as a one-dimensional representation of the pupil size across time. Although the data was recorded binocularly, the human coding used only the right eye.

The SMI HiSpeed 1250 eye-tracker ranked well on both accuracy and precision (See Table 2).

The present dataset

The eye tracking database

The eye tracking database from our group is fully described in Griffith et al. (2020) and is labelled "GazeBase". All details regarding the overall design of the database, subject recruitment, tasks and stimuli descriptions, calibration efforts, and eye tracking equipment are presented there. There were 9 temporally distinct "rounds" over a period of 37 months, and round 1 had the largest sample. This report only includes 19 subjects[10] randomly chosen from 322 subjects available from round 1. Briefly, subjects were initially recruited from the undergraduate student population at Texas State University through email and targeted in-class announcements. A total of 322 subjects (151-F, 171-M) were included in GazeBase (only 19 of which were used in the present study, see below.). Subjects completed two sessions of recording (median 19 min. apart) for each round of collection. Each session consisted of multiple tasks. The only task employed in the present evaluation was the text reading task. Therefore, this dataset will henceforth be labelled as the GazeBaseR dataset. Each subject was asked to read, silently, an identical pair of quatrains from the famous nonsense poem, "Hunting for a Snark", written by Lewis Carroll (written from 1874-1876). The text was displayed in Times New Roman 20 pt. size bold font and was single-spaced. The mean letter interval for each piece of text was approximately 0.50 degree of the visual angle. The height of the line of the text was 0.92 degrees of visual angle. Monocular (left) eye movements were captured at a 1,000 Hz sampling rate using an EyeLink 1000 eye tracker (SR Research, Ottawa, Ontario, Canada).

Eye movement recording

On each visit to our laboratory, subjects were studied twice (Sessions 1 and 2, for the present study, subjects were chosen randomly from either session), approximately 20 minutes apart. The subjects were seated 55 cm in front of a computer monitor with their heads resting on a chin/head rest. The monitor subtended +/- 23.3 degrees of visual angle in the horizontal direction, 11.7 degrees to the top and 18.5 degrees to the bottom. The mean spatial accuracy for our device was 0.71 (Friedman et al., 2021b), and the precision was 0.035 (Friedman et al., 2021a). The EyeLink 1000 eye-tracker was relatively highly ranked (Table 2). For further specifications, see EyeLink 1000 User Manual[11]. The sampling rate for our data was 1000 Hz. Data were calibrated using procedures provided by the manufacturer. The EyeLink 1000 transformed the raw records into gaze position data, in visual angle units, using the manufacturer-supplied calibration routine. If the EyeLink 1000 could not acquire a signal, as during blinks, NaN (not a number) was returned. Only the first 26 seconds of recordings for each subject were chosen for this dataset, because one of

[9] A few recordings in this dataset were mistakenly recorded at 200 Hz Friedman (2020)

[10] We started with 20 but dropped one very low quality recording.

[11] http://sr-research.jp/support/EyeLink%201000%20User%20Manual%201.5.0.pdf

the subjects chosen randomly finished reading the poem in approximately 26 seconds, and we wanted the same amount of data from each subject to be represented in each recording.

The training process

Here we describe the eye movement experience of the 3 raters in the present dataset:

All three raters attended a graduate-level course entitled "Human-Computer Interaction", taught by Dr. Komogortsev, which included substantial discussion of eye-tracking concepts. During that class the students became familiar with the basic concepts of eye movements and classic algorithms used for eye movement classification.

Rater 1: He was involved in an eye movement classification related project for a couple of months as a part-time activity. This project involved reading of the eye movement classification literature and development of a neural network for eye movement classification. He was never involved in the manual labeling of eye movement events.

Rater 2: He had used an eye tracker to collect data for his undergraduate thesis. He had also read a key reference text for eye movement reports (Holmqvist et al., 2011). He had never manually marked the start and end of any eye movement event.

Rater 3: He had never manually marked the start and end of any eye movement event.

Training Leader: The first author led the entire training effort. He has been studying eye movements, on and off, for more than 25 years. He has had considerable experience classifying and studying eye movements with 17 published papers on the topic. Most of these papers concerned smooth pursuit performance in psychiatric patient populations. In addition, the first author has published an extensively modified version of the widely cited algorithm of Nyström and Holmqvist (2010), (Friedman et al., 2018). Although he did rate the final dataset presented here, we do not present his ratings. What we are interested in contrasting is the inter-rater agreement between the 3 initially naive but extensively trained raters with the experts from the Hooge et al. (2018) and Larsson et al. (2013) studies. In the final analysis, the training leader had lower agreement with other raters than rater 1 but was comparable to raters 2 and 3.

There were seven iterative training rounds over a period of 4.6 months. From one to five training recordings were classified by each rater for each round. There were 20 consensus discussions, for a total time of 42.8 hours (average consensus discussion length was 2.14 hours). The scoring rules and a description of the interface we employed for classification can be found at https://digital.library.txstate.edu/handle/10877/13373.

All 3 raters classified all 26 seconds of all 19 datasets.

Preparing the eye movement recordings for analysis of agreement

Agreement statistics were analyzed between a single "ground truth" ("GT") rater and a comparison ("CMP") rater. All possible pairings of GTs and CMPs were analyzed. The first step was to create an array with 2 columns, and one row per sample, containing the coding for the GT rater (column 1) and the CMP rater (column 2). Fixations were coded as "1", saccades as "2" and PSEs[12] as "3". All other classified events (blinks, unclassified data, forms of noise and artifact) were also coded. We wanted to analyze only sections of the dataset where both raters classified either fixations, saccades or PSEs. If we had not done this, in certain cases, we would be comparing classification of good signal with non-classification due to the absence of classifiable signal (as during blinks and artefact). This would create classification errors in which there was not a true misclassification, but rather a classification versus a unclassifiable portion of a signal. Each such contiguous stretch was cut out of the entire recording and was referred to as a "snippet" Across all rater combinations and subjects, there was a minimum of 1 snippet and a maximum of 15 snippets (median = 4 snippets). For all 3 datasets, snippets shorter than 300 msec were discarded. Each snippet was analyzed separately, and agreement statistics were accumulated across snippets for each GT-CMP pair for each subject. For each snippet, there was a binarized version for fixation, saccades and PSEs. In the binarized version for fixation, fixation samples were labelled as 1 and non-fixation periods were labelled as 0. There were also binarized versions for saccades and PSEs. These binarized snippets were then assessed for sample-level and event-level agreement statistics.

Sample-level agreement statistics

To illustrate how this was done, consider the $2X2$ contingency table below (Table 3). Consider filling this table for a saccade binarized snippet. Every sample where both GT and CMP classified a saccade, that was a true positive (TP). For every sample where both GT and CMP classified a non-saccade, that was a true negative (TN). A sample where GT classified a non-saccade and CMP classified a saccade was considered a false positive (FP). A

[12] In this paper, we use the term post-saccadic event (PSE) rather than post-saccadic oscillation (PSO), since many saccades are followed by a single signal change without evidence of an oscillation. All signal changes that occur after a saccade ends and fixation resumes are considered PSEs.

Table 3 Contingency Table

	CMP=1	CMP=0
GT=1	TP(1,1)	FN(1,0)
GT=0	FP(0,1)	TN(0,0)

sample where GT classified a saccade and CMP classified a non-saccade was considered a false negative (FN). All the data from all the snippets for each subject for each pair of raters was accumulated in a table like Table 3, and Cohen's κ was computed.

Event level agreement statistics

The ultimate measure of event level agreement is the harmonic mean of "precision" and "recall" and is expressed as an F1-score (Powers, 2020). Typically, when referring to event-related analyses, we refer to true positives as "hits", true negatives as "misses" and false positives as "false alarms", abbreviated as "FA". For the table below, the symbol "#" signifies "number of". Also, during event-related analysis, true negatives are ignored. With marginal sums added, Table 3 above becomes Table 4 below:

"Precision" (also called positive predictive value) is the fraction of all events labelled as true (#hits+#FA) that are in fact true, or #hits/(#hits+#FA). "Recall" (also known as sensitivity) is the fraction of all events actually true (#hits+#misses) that were detected as true (#hits/(#hits+#misses).

Fig. 1 illustrates the event-level agreement analysis. The goal here was to assess the agreement between rater pairs at the event level (a contiguous fixation, saccade or PSE block of samples) in terms of an F1-score (Powers, 2020). Let's consider the measurement of saccade event-level agreement for an single snippet. The process proceeds as follows: We consider each GT-rated saccade event one at a time sequentially. For a given GT-rated saccade, there could be a number of different events (saccades or non-saccades) in the CMP datastream that overlaps with the GT identified event for at least 1 sample. One possibility was that the CMP scored a saccade at exactly the same point in time (start and end) as the GT rater. Another was that the CMP rater scored a saccade which started either before or after the GT rater or ended before or after the GT rater. Another was that the CMP rater scored no saccade at all. There are numerous other possibilities. For each GT-scored saccade event, we created an array of CMP events (CMP array) that intersected at least one sample with the GT saccade under analysis. Let's imagine that the CMP rater scored a saccade that started later than the GT saccade and ended earlier than the GT saccade. There would be 3 events in our array. The first would be the non-saccade period before the CMP-saccade, the second would be the CMP-rated saccade, and the third would be the CMP-rated non-saccade after the CMP-rated saccade. By processing these arrays we decide if the match was a hit or a miss. To do this, we calculate the intersection-over-union (IOU) ratio (also known as the Jaccard index Jaccard (1901), see also Startsev et al. (2019a)) between the GT-rated saccade and each event in the CMP array. We then sort our array of CMP events in descending order by the IOU ratio. If the first CMP event has an IOU ratio > 0.5 and it was classified as a saccade, we have a hit[13]. If the first event in our array has an IOU ratio > 0.5 and it was not classified as a saccade we have a miss. If none of the events in our array has an $IOU > 0.5$ we have a miss. If a CMP event in our array has already been used in a prior match, we have a miss. After all hits and misses have been found, any CMP event which was not matched to a GT event was considered as a false alarm. Hits, misses and false alarms were accumulated across snippets for each recording and each GT-CMP rater pair. For each recording and rater pair, we calculated the F1 score (Hooge et al., 2018; Powers, 2020; Rijsbergen, 1979) which was:

$$F1 = (2 * \#Hits)/(2 * \#\ddot{H}its + \#Misses + \#FA) \quad (1)$$

where "FA" means false alarms.

Including studies with a reasonable number of GT events

We use the term "study" to refer to a particular subject, rated by a single GT-CMP pair and for a single event type (fixation, saccade or PSE). We did not want to include any study which did not have enough GT events to produce a reliable and meaningful assessment. Therefore, we only calculated agreement level statistics on studies with 20 or more GT events.

Table 4 Contingency Table With Marginal Values

	CMP=1	CMP=0	
GT=1	#Hits	#Misses	#Hits+#Misses
GT=0	#FA		
	#Hits + #FA		

[13]The IOU threshold of 0.5 percent means that there must be at least 50% overlap between two events to conclude that the event classification agrees. We used this as did Startsev et al. (2019a). According to Startsev et al. (2019a), this is the lowest threshold that ensures that no two detected events can be candidate matches for a single ground truth event. Additionally, if two events have the same duration, their relative shift can be no more than one-third of their duration.

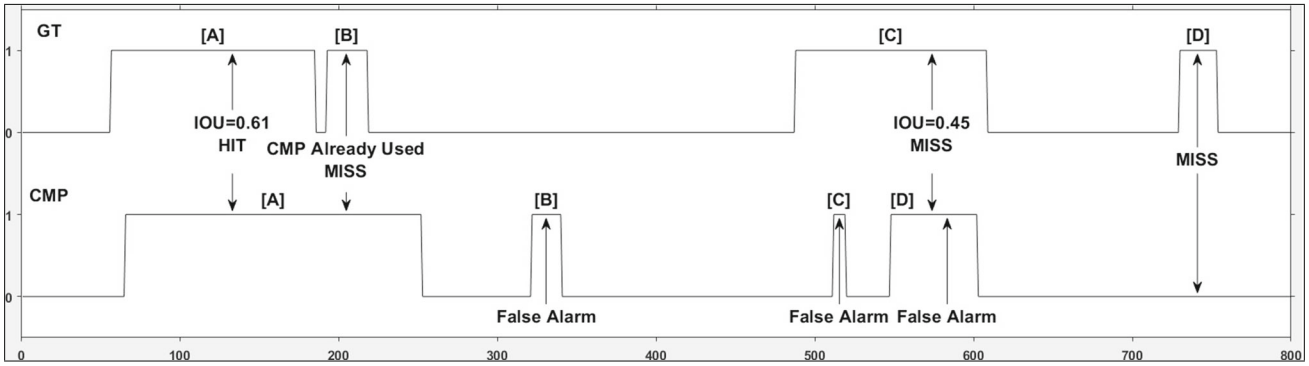

Fig. 1 Illustration of the analysis of event-level agreement. The top line represents the event classification for the GT rater, and the bottom line represents the event classification for the CMP rater. Events occur at level 1, and non-events occur at level 0. Each GT event and each CMP is labelled either '[A]','[B]','[C]' or '[D]'. The x-axis is time, in milliseconds. GT event [A] has an intersection-over-union (IOU) ratio of 0.61 with CMP event [A]. Since this IOU is greater than 0.5, this match is considered a hit. GT event B cannot also match with CMP event [A] because CMP event [A] is already matched. Therefore, this match is considered a miss. CMP event [B] is unmatched to any GT event, and so it is considered a false alarm. CMP event C is also not matched to any GT event so it is also classified as a false alarm. The IOU for GT event [C] and CMP event [D] is less than 0.5 so this match is considered a miss. It is also considered a false alarm, since CMP event [D] is not matched to any GT event. GT event [D] has no matching CMP event and so this is considered a miss

Statistical analysis

Differences between datasets for sample-level κ and event-level F1-scores were tested with the Kruskal-Wallis test. A non-parametric test was chosen after visual inspection of Figs. 2 to 7 revealed mostly non-normal distributions. It tests the null hypothesis that the distribution of the dependent variable (κ or F1) is the same across datasets. When appropriate, these tests were followed up with post-hoc comparisons. Multiple comparisons were controlled using the Bonferroni method.

Results

Included studies

We use the term "study" to refer to a single comparison between two raters (GT and CMP) for a particular subject for a particular event type. With 3 raters, there were:

$$[(3 \cdot 3) - 3) = 6]$$

rater combinations. So, for the present data, with 19 subjects, there were 19*6 or 114 studies (per event type). As noted above, we did not calculate agreement statistics (sample- or event-level) on studies which did not have at least 20 GT events. Table 5 provides an accounting of the number of studies of each type, the number of rejected studies and the percentage of rejected studies.

Sample-level Agreement Statistics

The dot markers in Figs. 2 to 4 are median sample-level scores for each GT against each available CMP rater. The numbers plotted sideways in these plots are the medians across raters (medians of medians). For this paragraph and the next 5, when we refer to median, we are referring to the medians across raters. The sample-level κ for fixations are

Table 5 Studies rejected based on too few events

Dataset	Event Typet	Total Studies	Number Rejected	Percent Rejected
GazeBaseR	Fixation	114	0	0.0
GazeBaseR	Saccades	114	0	0.0
GazeBaseR	PSEs	114	40	35.1
Lund	Fixation	12	1	8.3
Lund	Saccades	12	1	16.7
Lund	PSEs	12	2	16.7
Hooge	Fixation	1320	0	0.0

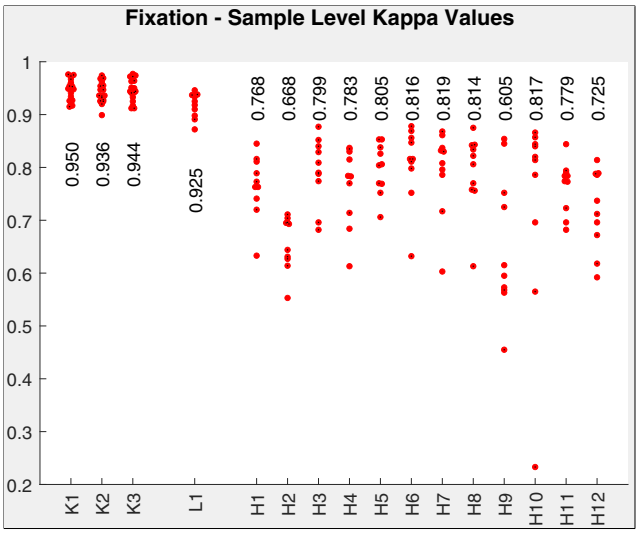

Fig. 2 Sample-level κ for the classification of fixations. K-1...K3 refer to the 3 raters in the present dataset ("GazeBaseR dataset"). L-1 refers to rater MN from the Lund dataset. H-1...H12 refer to the twelve raters in the Hooge et al. (2018) report. These codes refer to the ground truth rater. Each plot contains N recording points (GazeBaseR = 19, Lund = 12, Hooge = 10). For each subject x, the κ plotted is the median κ for all other raters versus the ground truth rater

illustrated in Fig. 2. All of the raters from the present dataset have median sample-level κ values that were between 0.944 and 0.950. These were higher than the sample-level κ for the Lund dataset (median = 0.925), and also substantially higher than the 12 raters from Hooge et al. (2018) (ranged from 0.605 to 0.819). The median κ across all recordings within a dataset are in Table 6. There were statistically significant differences across datasets (Table 6). Post-hoc comparisons revealed that the raters from either the Lund or GazeBaseR datasets outperformed the raters in the Hooge dataset (Table 6). The GazeBaseR raters outperformed the raters for the Lund dataset (Table 6).

The median sample-level κ for saccades for the present dataset and the Lund dataset are illustrated in Fig. 3. For the raters from the present dataset, the median sample-level κ ranged from 0.969 to 0.972. They were all higher than the sample-level κ for the Lund dataset (median = 0.941). The medians across recordings within a dataset are in Table 6. There were statistically significant differences. Raters from the GazeBaseR datasets outperformed the raters in the Lund dataset (Table 6).

The median sample-level κ for PSEs for the present dataset and the Lund dataset are illustrated in Fig. 4. The κ for the raters in this dataset range from 0.800 to 0.813. The medians across recordings within a dataset are in Table 6. These medians were not statistically different (Table 6) across datasets.

Event level agreement statistics

Event-level results for fixation are illustrated in Fig. 5. For the present dataset, the range of median F1-scores was from 0.993 to 0.995. These were very similar to the F1-score for the Lund dataset (median = 1.00). They were in every case higher than the F1-scores of the 12 raters from Hooge et al. (2018) (range = 0.849 to 0.951). The medians across recordings within a dataset are in Table 6. There were statistically significant differences across datasets (Table 6). Post-hoc comparisons revealed that the raters from either the Lund or GazeBaseR datasets outperformed the raters in the Hooge dataset (Table 6). The GazeBaseR and Lund raters were not statistically different (Table 6).

Event-level F1-scores for saccades are illustrated in Fig. 6. For the present dataset, the range of median F1-scores were from 0.991 to 0.995. These were slightly lower than the median F1-score for the Lund dataset (1.00). The medians across recordings within a dataset are in Table 6. There were no statistically significant differences.

Table 6 Median κ and F1-scores across datasets for All Event Types

Event Type	Metric	GazeBaseR Dataset	Lund Dataset	Hooge DataSet	Omnibus p-value	K-L Post-hoc p-value	K-H Post-hoc p-value	L-H Post-hoc p-value
FIX	κ	0.947	0.925	0.764	< 0.001	< 0.002	< 0.001	< 0.001
SAC	κ	0.970	0.941		< 0.005			
PSE	κ	0.807	0.768		***ns***			
FIX	F1-score	1.000	1.000	0.930	< 0.001	***ns***	< 0.001	< 0.001
SAC	F1-score	0.994	1.000		***ns***			
PSE	F1-score	0.792	0.804		***ns***			

K=GazeBaseR, L=Lund, H=Hooge. FIX = fixation, SAC = saccade, ***ns*** = not statistically significant. Omnibus test was a Kruskal Wallis test. For fixation, with 3 datasets compared, df = 2. For saccades and PSEs, df = 1. Post-hoc tests were Mann-Whitney tests. These p-values would pass any form of correction for multiple comparisons. Since our focus was on the initially-naive raters, rater 4 was excluded from these analyses

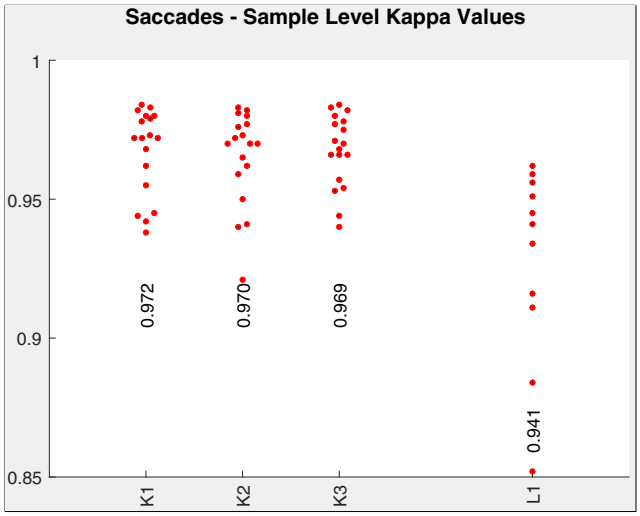

Fig. 3 Sample-level κ for the classification of saccades. See caption for **Fig. 1** for details. Raters in Hooge et al. (2018) did not classify saccades

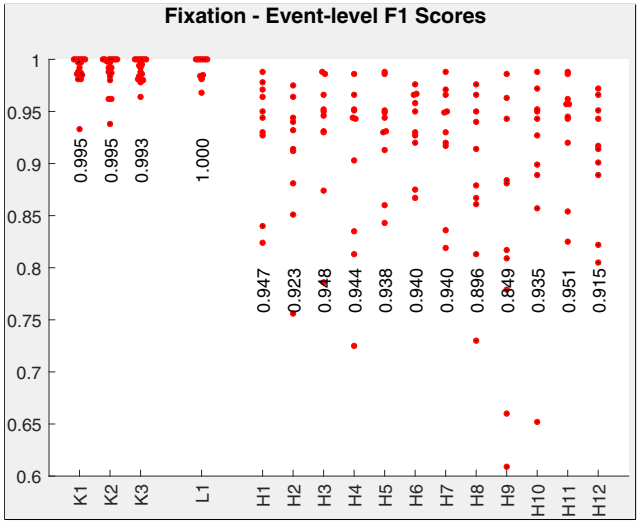

Fig. 5 Event-level F1 scores for the classification of fixations. See caption for **Fig. 1** for details

Event-level F1-scores for PSEs are illustrated in Fig. 7. For the present dataset, the range of median F1-scores were from 0.777 to 0.804. These were slightly lower than the F1-score for the Lund dataset (0.804). The medians across recordings within a dataset are in Table 6. There were no statistically significant differences.

Discussion

We wanted to compare our agreement performance with that from other raters on other datasets. We want to make it perfectly clear that we are fully aware that these other datasets recorded eye movements with different devices, were collected while subjects viewed different stimuli, used recordings from different subjects and employed different raters. Nonetheless, we believe these comparisons have some general value and provide some insight.

The main finding of the present report was that inter-rater agreement was lowest for the dataset based on low-quality data and expert but untrained raters (Hooge), was much better in a dataset based on high-quality data and two exceptionally qualified raters (LUND), but was the best in our dataset based on high quality data, and employing initially naive raters who were extensively trained. For fixation, both the GazeBaseR dataset and the LUND dataset had statistically higher inter-rater agreement than

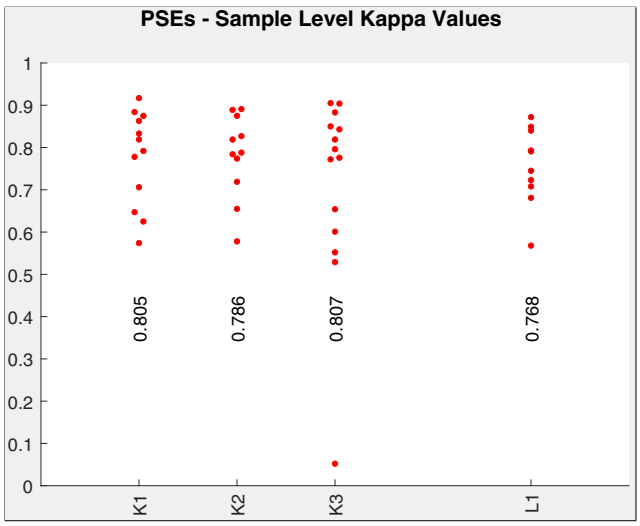

Fig. 4 Sample-level κ for the classification of PSEs. See caption for **Fig. 1** for details. Raters in Hooge et al. (2018) did not classify PSEs

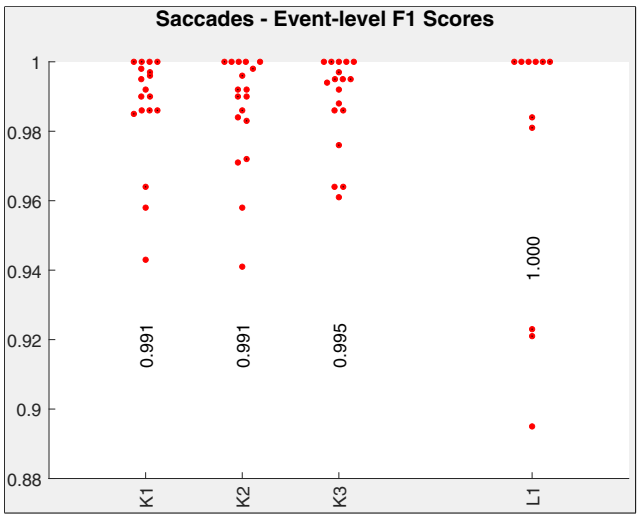

Fig. 6 Event level F1 scores for the classification of saccades. See caption for **Fig. 1** for details. Raters in Hooge et al. (2018) did not classify saccades

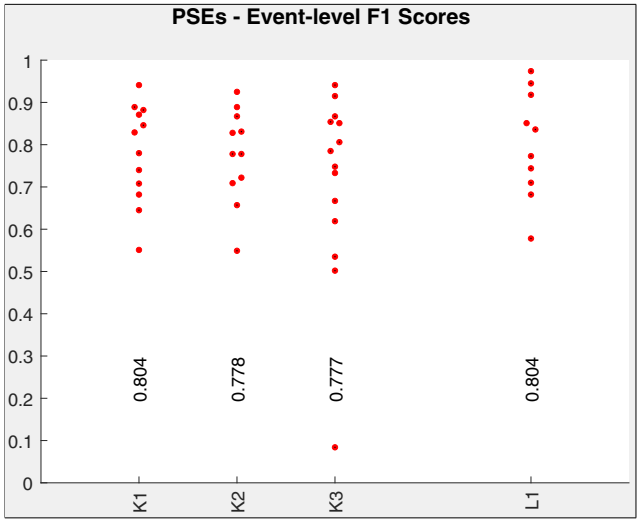

Fig. 7 Event level F1 scores for the classification of PSEs. See caption for **Fig. 1** for details. Raters in Hooge et al. (2018) did not classify PSEs

the Hooge dataset at both the sample and the event level. Note that the agreement statistics for the Hooge et al. (2018) study would have, in all likelihood, been even lower if we had included their infant data. Our classification of the GazeBaseR dataset statistically outperformed the expert classification from the LUND dataset for sample-level agreement for fixation and saccades. Since our classifications outperformed the LUND dataset for fixation and saccade sample-level agreement but not event-level agreement, it is reasonable to hypothesize that our initially naive raters were more consistent (after training) with the timing (onset and offset) of events, and not more consistent with identifying the presence of an event.

We cannot state, based on scientific evidence, which of the dataset characteristics (quality of eye-tracking signal, presence /absence of head restraint, subject type, rater training present or absent) had the greatest effect. Separate evaluations of each factor would be required to determine that. For example, studies where data of different eye-tracking signal quality was classified by the same set raters would address the role of eye tracking signal quality. Studies comparing a set of expert untrained raters to initially naive and trained raters on the same data would address the issue of rater type. However, given the very large time and personnel resources required for such studies, these issues may not be resolved for some time.

High inter-rater agreement does not directly address the issue of classification accuracy. Raters may agree but all be inaccurate. We are not sure how to test classification accuracy. However, in the present dataset, unlike prior reports, we provide detailed scoring rules which others can use to judge accuracy and compare our definitions to those employed in other laboratories.

Of course, not every study can use a high quality eye-tracker, head restraint and healthy adult subjects. The Hooge et al. (2018) paper, and the data presented here give some sense of what is lost in terms of inter-rater classification agreement when different approaches for rater training and different eye tracking signal quality levels are considered.

When manually classifying basic eye movements, we consider high inter-rater reliability an important indication of a meaningful standardized classification that can be subsequently employed as comparisons to results from new automatic classification methods and also to train machine learning techniques to provide accurate classification results.

We were able to achieve high inter-rater reliability by employing a detailed iterative approach for rater training. We understand that such an approach might be very time consuming, especially in cases of captured eye tracking signal of various quality levels and a variety of underlying eye movement types. However, we strongly believe that such an approach to manual ground truth labeling of basic eye movement types will provide the most accurate and thus useful classification results.

Acknowledgements This work was funded by grant from the NSF (1714623) (PI: Oleg Komogortsev).

Data Availability We are currently further analyzing the results of this study. Once we are finished with this, the dataset, including recordings and manual classification results will be made available at https://digital.library.txstate.edu/handle/10877/13373.

Declarations

Conflict of Interests We have no conflict of interest.

References

Agtzidis, I., Startsev, M., & Dorr, M. (2020). Two hours in Hollywood: A manually annotated ground truth data set of eye movements during movie clip watching. Journal of Eye Movement Research, 13(4).

Andersson, R., Larsson, L., Holmqvist, K., Stridh, M., & Nyström, M. (2017). One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behavior Research Methods*, *49*(2), 616–637.

Dar, A.H., Wagner, A.S., & Hanke, M. (2021a). REMoDNaV: robust eye-movement classification for dynamic stimulation. *Behavior Research Methods*, *53*(1), 399–414.

Dar, A.H., Wagner, A.S., & Hanke, M. (2021b). REMoDNaV: robust eye-movement classification for dynamic stimulation. *Behavior Research Methods*, *53*(1), 399–414.

Friedman, L. (2020). Brief communication: Three errors and two problems in a recent paper: gazeNet: End-to-end eye-movement event detection with deep neural networks (Zemblys, Niehorster, and Holmqvist, 2019). *Behavior Research Methods*, *52*(4), 1671–1680.

Friedman, L., Hanson, T., & Komogortsev, O.V. (2021a). Multimodality during fixation – Part II: Evidence for multimodality in spatial precision-related distributions and impact on precision estimates. Journal of Eye Movement Research, 14(3).

Friedman, L., Lohr, D., Hanson, T., & Komogortsev, O.V. (2021b). Angular Offset Distributions During Fixation Are, More Often Than Not, Multimodal. Journal of Eye Movement Research,14(3).

Friedman, L., Rigas, I., Abdulin, E., & Komogortsev, O.V. (2018). A novel evaluation of two related and two independent algorithms for eye movement classification during reading. *Behavior Research Methods*, *50*(4), 1374–1397.

Friedman, L. (2020). A Re-Examination of the Evidence used by Hooge et al (2018) "Is human classification by experienced untrained observers a gold standard in fixation detection?". arXiv:2001.07701.

Fuhl, W., & Kasneci, E. (2021). A Multimodal Eye Movement Dataset and a Multimodal Eye Movement Segmentation Analysis. arXiv:2101.04318.

Griffith, H., Lohr, D., Abdulin, E., & Komogortsev, O. (2020). GazeBase: A Large-Scale, Multi-Stimulus, Longitudinal Eye Movement Dataset. arXiv:2009.06171.

Holmqvist, K. (2017). Common predictors of accuracy, precision and data loss in 12 eye-trackers (available on researchgate). In *The 7th Scandinavian Workshop on Eye Tracking*.

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

Hooge, I.TC., Niehorster, D.C., Nyström, M., Andersson, R., & Hessels, R.S. (2018). Is human classification by experienced untrained observers a gold standard in fixation detection?. *Behavior Research Methods*, *50*(5), 1864–1881.

Hooge, I.TC., Niehorster, D.C., Nyström, M., Andersson, R., & Hessels, R.S. (2021). Correction to: "Is human classification by experienced untrained observers a gold standard in fixation detection?". Behavior Research Methods.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles*, *37*, 547?579.

Jongerius, C., Callemein, T., Goedem?, T., Van Beeck, K., Romijn, J.A., Smets, E.MA., & Hillen, M.A. (2021). Eye-tracking glasses in face-to-face interactions: Manual versus automated assessment of areas-of-interest. *Behavior Research Methods*, *53*(5), 2037–2048.

Korda, A.I., Asvestas, P.A., Matsopoulos, G.K., Ventouras, E.M., & Smyrnis, N.P. (2015). Automatic identification of oculomotor behavior using pattern recognition techniques. *Computers in Biology and Medicine*, *60*, 151–162.

Kothari, R., Yang, Z., Kanan, C., Bailey, R., Pelz, J.B., & Diaz, G.J. (2020). Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Science Reports*, *10*(1), 2539.

Larsson, L., Nyström, M., & Stridh, M. (2013). Detection of saccades and postsaccadic oscillations in the presence of smooth pursuit. *IEEE Transactions on Biomedical Engineering*, *60*(9), 2484–2493.

Niehorster, D.C., Cornelissen, T.HW., Holmqvist, K., Hooge, I.TC., & Hessels, R.S. (2018). What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods*, *50*(1), 213–227.

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*(1), 188–204.

Powers, D.MW. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv:2010.16061.

Rijsbergen, C.JV. (1979). Information retrieval, 2nd ed. Butterworths.

Startsev, M., Agtzidis, I., & Dorr, M. (2019a). 1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits. *Behavior Research Methods*, *51*(2), 556–572.

Startsev, M., Agtzidis, I., & Dorr, M. (2019b). Characterizing and automatically detecting smooth pursuit in a large-scale ground-truth data set of dynamic natural scenes. *Journal of Vision*, *19*(14), 10.

Startsev, M., Agtzidis, I., & Dorr, M. (2019c). Characterizing and automatically detecting smooth pursuit in a large-scale ground-truth data set of dynamic natural scenes. *Journal of Vision*, *19*(14), 10.

Stuart, S., Hunt, D., Nell, J., Godfrey, A., Hausdorff, J.M., Rochester, L., & Alcock, L. (2018). Do you see what I see? Mobile eye-tracker contextual analysis and inter-rater reliability. *Medical and Biological Engineering and Computing*, *56*(2), 289–296.

Vargas-Cuentas, N.I., Roman-Gonzalez, A., Gilman, R.H., Barrientos, F., Ting, J., Hidalgo, D., ..., Zimic, M. (2017). Developing an eye-tracking algorithm as a potential tool for early diagnosis of autism spectrum disorder in children. *PloS One*, *12*(11), e0188826.

Venker, C.E., Pomper, R., Mahr, T., Edwards, J., Saffran, J., & Ellis Weismer, S. (2020). Comparing Automatic Eye Tracking and Manual Gaze Coding Methods in Young Children with Autism Spectrum Disorder. *Autism Research*, *13*(2), 271–283.

Wadehn, F., Mack, D.J., Weber, T., & Loeliger, H.A. (2018). Estimation of Neural Inputs and Detection of Saccades and Smooth Pursuit Eye Movements by Sparse Bayesian Learning. *Annu Int Conf IEEE Eng Med Biol Soc*, *2018*, 2619–2622.

Wadehn, F., Mack, D.J., Weber, T., & Loeliger, H.A. (2018). Estimation of Neural Inputs and Detection of Saccades and Smooth Pursuit Eye Movements by Sparse Bayesian Learning. *Annu Int Conf IEEE Eng Med Biol Soc*, *2018*, 2619–2622.

Wang, D., Mulvey, F.B., Pelz, J.B., & Holmqvist, K. (2017). A study of artificial eyes for the measurement of precision in eye-trackers. *Behavior Research Methods*, *49*(3), 947–959. https://doi.org/10.3758/s13428-016-0755-8.

Zemblys, R., Niehorster, D.C., & Holmqvist, K. (2019). gazeNet: End-to-end eye-movement event detection with deep neural networks. *Behavior Research Methods*, *51*(2), 840–864.

Zemblys, R., Niehorster, D.C., Komogortsev, O., & Holmqvist, K. (2018). Using machine learning to detect events in eye-tracking data. *Behavior Research Methods*, *50*(1), 160–181.