Reporting Eye-Tracking Data Quality: Towards a New Standard

Deborah N. Jakobi*
jakobi@cl.uzh.ch
University of Zurich
Zurich, Switzerland

Daniel G. Krakowczyk*
daniel.krakowczyk@uni-potsdam.de
University of Potsdam
Potsdam, Germany

Lena A. Jäger
jaeger@cl.uzh.ch
University of Zurich
Zurich, Switzerland
University of Potsdam
Potsdam, Germany

ABSTRACT

Eye-tracking datasets are often shared in the format used by their creators for their original analyses, usually resulting in the exclusion of data considered irrelevant to the primary purpose. In order to increase re-usability of existing eye-tracking datasets for more diverse and initially not considered use cases, this work advocates a new approach of sharing eye-tracking data. Instead of publishing filtered and pre-processed datasets, the eye-tracking data at all pre-processing stages should be published together with data quality reports. In order to transparently report data quality and enable cross-dataset comparisons, we develop data quality reporting standards and metrics that can be automatically applied to a dataset, and integrate them into the open-source Python package *pymovements* (https://github.com/aeye-lab/pymovements).

CCS CONCEPTS

• **Applied computing** → **Psychology**.

KEYWORDS

Eye-tracking, data quality, reporting standards, reproducibility, transparency

ACM Reference Format:
Deborah N. Jakobi, Daniel G. Krakowczyk, and Lena A. Jäger. 2024. Reporting Eye-Tracking Data Quality: Towards a New Standard. In *2024 Symposium on Eye Tracking Research and Applications (ETRA '24), June 04–07, 2024, Glasgow, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3649902.3655658

1 INTRODUCTION

Eye-tracking-while-reading data has been used for a diverse range of purposes that range from addressing psycholinguistic research questions [Rayner and Pollatsek 2006] to improving downstream applications in Natural Language Processing (NLP) [Barrett et al. 2018; Deng et al. 2023] or biometric reader identification [Jäger et al. 2020]. While data that is considered to be bad quality (e.g., low calibration scores) is often discarded in psycholinguistic research, more application-oriented research often requires vast amounts of data to train machine learning models and in some cases specifically requires data of non-optimal quality, as this can help to improve the model's generalizability to use case where only low-quality data is available [Prasse et al. 2020]. Along the same lines, different use cases require data at different pre-processing stages. While some research questions typically focus on the analysis of aggregated reading measures on word- or sentence-level, other research uses sequence models to model scanpaths [Reich et al. 2022; von der Malsburg and Vasishth 2011], or models raw gaze data [Prasse et al. 2024]. To completely enable the re-usability of datasets for use cases that were not initially considered by the creators, *data at all pre-processing stages* needs to be made available. We propose a paradigmatic shift away from only publishing the filtered and pre-processed data that was used for the original data analysis towards a fully transparent and reproducible approach where no data is discarded. In particular, we propose to provide detailed metadata including data quality reports together with a fully transparent and reproducible pre-processing pipeline to enable the future users of the dataset to decide about their inclusion criteria and pre-processing.

Eye-tracking data quality refers to the quality of the positional data collected during eye-tracking; that is how well the raw gaze data samples represent the actual movements of the eyes [Holmqvist et al. 2011]. Often, the quality is described in terms of spatial accuracy, temporal and spatial resolution (sampling frequency and precision), and data loss, all of which capture different aspects of the relationship between the true and the measured gaze position [Hessels and Hooge 2019]. To date, there exist no concrete standards or metrics as to what is considered *high quality* eye-tracking data for a specific use case. Many different and diverse factors are known to influence data quality such as the sampling frequency, the calibration procedures or the suitability of the stimulus size. Both, properties of the experimental setup that influence data quality and known data quality metrics such as calibration accuracy, need to be reported such that re-users can make informed decisions on whether the data might be useful for their research.

The goal of our present research is to provide concrete suggestions and precise definitions of which properties of the setup and data quality measures to report, and we present a set of data quality metrics for dataset comparison. Following the FAIR principles of accessible, re-usable and interoperable data and code [Wilkinson et al. 2016] and to ensure comparability across devices and datasets, reproducibility, and transparency, we integrate our implementations into the Python package *pymovements* [Krakowczyk et al. 2023].

*Both authors contributed equally to this research.

ETRA '24, June 04–07, 2024, Glasgow, United Kingdom

ACM ISBN 979-8-4007-0607-3/24/06
https://doi.org/10.1145/3649902.3655658

2 RELATED WORK

A first line of research on eye-tracking data quality investigates what factors influence data quality. For example, Andersson et al. [2010] study how sampling frequency influences the detection of gaze events. A second line of research actively seeks solutions to improve data quality either at recording time or post-hoc. Examples of this line of research are Nyström et al. [2013], who study how calibration procedures can be improved, or Carr et al. [2022], who investigate how vertical calibration drift can be corrected post-hoc using different algorithms [Carr 2023]. A third line of research seeks to standardize metadata reports about the experimental setup and procedure, and/or data quality. To date, there is no clear standard regarding which aspects of the setup/procedure of eye-tracking data collections should be reported, nor is there consensus on how to assess and report data quality. Oftentimes, information about the experimental setup and procedure that impact data quality, such as specifics on the calibration procedure, is missing. This issue of a lack of transparent data quality reporting was raised by different researchers [Dunn et al. 2023; Hessels and Hooge 2019; Nyström et al. 2013]. A recent advancement in this area is the Python package *pymovements*, that is in line with the open-source standards of complete transparency and reproducibility in combination with standardized documentation [Krakowczyk et al. 2023]. The package allows users to select and apply different common eye-tracking preprocessing algorithms with freely chosen parameters but openly available source code which makes the pre-processing pipeline completely transparent and reproducible (if the parameters are shared by the users). In our work, we expand this package to include data quality reports that follow the same practices of transparency, reproducibility, open-source and standardized documentation.

3 REPORTING EYE-TRACKING DATA QUALITY

We provide suggestions what metadata to report together with concrete implementations as well as a set of specific data quality metrics to create more transparent and re-usable eye-tracking datasets. Currently, we focus on trial- and session-level data quality reports for reading research conducted with different EyeLink devices but the general approach can be applied to any type of eye-tracking recordings.

3.1 Publishing data at all pre-processing stages

While traditional reading research analyses eye movement data almost exclusively as aggregated word-level reading measures (such as a word's first-pass fixation duration), more recent research started to use non-aggregated fixation data (e.g., scanpath analyses [Reich et al. 2022; von der Malsburg and Vasishth 2011]) or even modelled the raw gaze samples [Jäger et al. 2020; Prasse et al. 2024]. Eye-tracking research can continuously benefit if the data at all preprocessing stages including raw gaze samples is published. Most importantly, the raw gaze samples needs to be available in order to compute metrics such as the data loss ratio. When publishing eye-tracking data, it should be kept in mind that privacy issues can arise for both raw and aggregated data since eye gaze data at different preprocessing stages have been shown to exhibit strong idiosyncracies that can be used for biometric reader or viewer identification [Holland and Komogortsev 2011; Makowski et al. 2020, 2019, 2021].

3.2 Reporting metadata

Each dataset should come with a set of standardized metadata on different levels (trial-, session-, and dataset-level, etc.). We therefore propose to automatically generate standardized metadata reports which we have integrated in *pymovements*. They include, for example, the following metadata: sampling rate, tracked eye(s), filter settings, date and time, total recording duration, eye tracker model and version information, and display resolution.

3.3 Reporting validation and calibration

Calibration and validation routines are included in the standard eye-tracking experiment procedure. Often, validations and/or recalibrations are performed multiple times throughout experiments which should be reported as validation scores can point towards decreasing data quality. Our implementation includes information about the number and timestamps of calibrations/validations performed, the average and maximum scores, and the error label given by the recording device for each validation, and the tracked eye, as well as session-level settings such as the number of calibration points.

3.4 Reporting data loss

Data loss refers to samples that were not recorded, for example, due to blinks or head movements. This means that the data contains less samples than there should be given the sampling frequency [Hessels and Hooge 2019]. The reports we implemented include start and stop timestamps of blinks detected by the EyeLink software, and the duration and number of samples for each blink, the overall data loss ratio including blinks and the data loss ratio caused by detected blinks. The difference between the two denotes samples lost due to unknown causes. A high blink ratio can further indicate that the participant felt uncomfortable.

3.5 Data quality metrics

While the above mentioned reports are independent from the presented stimulus, we propose to additionally report **stimulus dependent metrics** that can be used to estimate the statistical validity of eye-tracking-while-reading data. In particular, we propose to report the word-skipping-rate, the background dwell-time, and the ratio of line jumps over multiple lines in standardized ways for all datasets. Furthermore, we propose to create metrics that incorporate well-studied psycholinguistic effects (e.g., the effect of age on the reading speed or the word length effect). If the same metrics are applied across datasets it will be possible to determine whether the reading behaviour recorded from a certain subject is indicative of inattentive reading, bad calibration or other factors.

4 CONCLUSION & FUTURE WORK

This work contributes to setting the foundation of a continued effort to i) increase the re-usability of eye-tracking data, ii) increase the transparency of eye-tracking data quality, and iii) enable comparisons between different eye-tracking datasets all of which strive towards making eye-tracking datasets align more with the FAIR principles [Wilkinson et al. 2016]. We leave it to future research to extend the reporting standards and refine existing ones.

5 BROADER IMPACT

Reporting eye-tracking data quality can greatly improve any research in the field. While it is technically possible to obtain reports about, for example, blinks using the standard Data Viewer software to analyze EyeLink data files [SR Research Ltd. 2011], these reports are eye-tracker-specific, and the exact implementation depends therefore on the manufacturer. In order to standardize these reports and shift towards a higher degree of transparency, it is of great value to use an open-source package that can be adapted to whichever device was used. Moreover, by facilitating the sharing of experiment details and gaze data at different preprocessing stages, we aim to pave the way for more research into ensuring high data quality and defining data quality standards across the field. As a consequence of these endeavours, it will be unavoidable to re-discuss and possibly change the role of manufacturers and proprietary restrictions on algorithms and data formats, and develop robust metadata and quality reporting standards which do not rely on closed-source proprietary software.

ACKNOWLEDGMENTS

This work was partially funded by the Swiss National Science Foundation under grant 212276, by the German Federal Ministry of Education and Research under grant 01|S20043 and is based upon work from COST Action MultiplEYE, CA21131, supported by COST (European Cooperation in Science and Technology).

REFERENCES

Richard Andersson, Marcus Nyström, and Kenneth Holmqvist. 2010. Sampling frequency and eye-tracking measures: How speed affects durations, latencies, and more. *Journal of Eye Movement Research* 3, 3 (2010).

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. 302–312.

Jon W. Carr. 2023. eyekit: A lightweight Python package for doing open, transparent, reproducible science on reading behavior. https://github.com/jwcarr/eyekit

Jon W. Carr, Valentina N. Pescuma, Michele Furlan, Maria Ktori, and Davide Crepaldi. 2022. Algorithms for the automated correction of vertical drift in eye-tracking data. *Behavior Research Methods* 54 (2022), 287–310.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2023. Pretrained language models augmented with synthetic scanpaths for natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 6500–6507.

Matt J. Dunn, Robert G. Alexander, Onyekachukwu M. Amiebenomo, Gemma Arblaster, Denize Atan, Jonathan T. Erichsen, Ulrich Ettinger, Mario E. Giardini, Iain D. Gilchrist, Ruth Hamilton, Roy S. Hessels, Scott Hodgins, Ignace T. C. Hooge, Brooke S. Jackson, Helena Lee, Stephen L. Macknik, Susana Martinez-Conde, Lee Mcilreavy, Lisa M. Muratori, Diederick C. Niehorster, Marcus Nyström, Jorge Otero-Millan, Michael M. Schlüssel, Jay E. Self, Tarkeshwar Singh, Nikolaos Smyrnis, and Andreas Sprenger. 2023. Minimal reporting guideline for research involving eye tracking (2023 edition). *Behavior Research Methods* (2023).

Roy S. Hessels and Ignace T. C. Hooge. 2019. Eye tracking in developmental cognitive neuroscience – The good, the bad and the ugly. *Developmental Cognitive Neuroscience* 40, 100710 (2019).

Corey Holland and Oleg V. Komogortsev. 2011. Biometric identification via eye movement scanpaths in reading. In *2011 International Joint Conference on Biometrics (IJCB)*. 1–8.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, Joost van de Weijer, Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures.* Oxford University Press, Oxford, New York.

Lena A. Jäger, Silvia Makowski, Paul Prasse, Sascha Liehr, Maximilian Seidler, and Tobias Scheffer. 2020. Deep Eyedentification: Biometric identification using micro-movements of the eye. In *Machine Learning and Knowledge Discovery in Databases*, Ulf Brefeld, Elisa Fromont, Andreas Hotho, Arno Knobbe, Marloes Maathuis, and Céline Robardet (Eds.). Springer, Cham, Switzerland, 299–314.

Daniel G. Krakowczyk, David R. Reich, Jakob Chwastek, Deborah N. Jakobi, Paul Prasse, Assunta Süss, Oleksii Turuta, Paweł Kasprowski, and Lena A. Jäger. 2023. pymovements: A Python package for eye movement data processing. In *Proceedings of the 2023 ACM Symposium on Eye Tracking Research and Applications (ETRA '23)*. 53.

Silvia Makowski, Lena A. Jäger, Paul Prasse, and Tobias Scheffer. 2020. Biometric identification and presentation-attack detection using micro- and macro-movements of the eyes. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB 2022)*. IEEE.

Silvia Makowski, Lena A. Jäger, Ahmed Abdelwahab, Niels Landwehr, and Tobias Scheffer. 2019. A discriminative model for identifying readers and assessing text comprehension from eye movements. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I.* Springer-Verlag, Berlin, Heidelberg, 209–225.

Silvia Makowski, Paul Prasse, David R. Reich, Daniel Krakowczyk, Lena A. Jäger, and Tobias Scheffer. 2021. DeepEyedentificationLive: Oculomotoric biometric identification and presentation-attack detection using deep neural networks. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 4 (2021), 506–518.

Marcus Nyström, Richard Andersson, Kenneth Holmqvist, and Joost van de Weijer. 2013. The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods* 45 (2013), 272–288.

Paul Prasse, Lena A. Jäger, Silvia Makowski, Moritz Feuerpfeil, and Tobias Scheffer. 2020. On the relationship between eye tracking resolution and performance of oculomotoric biometric identification. *Procedia Computer Science* 176 (2020), 2088–2097.

Paul Prasse, David R. Reich, Silvia Makowski, Tobias Scheffer, and Lena A. Jäger. 2024. Improving cognitive-state analysis from eye gaze with synthetic eye-movement data. *Computers & Graphics* 119 (2024).

Keith Rayner and Alexander Pollatsek. 2006. Eye-movement control in reading. In *Handbook of Psycholinguistics* (2nd ed.), Matthew J. Traxler and Morton A. Gernsbacher (Eds.). Academic Press, London, UK, Chapter 16, 613–657.

David Robert Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, and Lena A. Jäger. 2022. Inferring native and non-native human reading comprehension and subjective text difficulty from scanpaths in reading. In *2022 Symposium on Eye Tracking Research and Applications (ETRA '22)*. Association for Computing Machinery, 1–8.

SR Research Ltd. 2011. *EyeLink Data Viewer user's manual.* Mississauga, Canada. Document version 1.11.1.

Titus von der Malsburg and Shravan Vasishth. 2011. What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language* 65, 2 (2011), 109–127.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, and et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 160018.