# Investigating pupillometry as a reliable measure of individual's listening effort

MIHAELA-BEATRICE NEAGU[1,*], TORSTEN DAU[1], PETTERI HYVÄRINEN[1], PER BÆKGAARD[2], THOMAS LUNNER[3,4], DOROTHEA WENDT[1,3]

[1] *Hearing Systems, Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark*

[2] *Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark*

[3] *Eriksholm Research Center Denmark, Snekkersten, Denmark*

[4] *Department of Electrical Engineering, Linköping University, Linköping, Sweden*

Pupillometry as a tool indicating listening effort has been extensively analyzed on a group level, but less is known about how reliable pupil dilation is as an indicator of an individual's listening effort. The aim of this study was to investigate the reliability of the pupil dilation measured during a speech-in-noise task as an indicator of an individual's listening effort. The pupil dilation of 27 normal-hearing (NH) and 24 hearing-impaired (HI) participants was recorded while they performed a speech-in-noise test on two different days. Measures of intraclass correlation coefficient (ICC) absolute agreement were considered in the analysis. The ICC was applied to the peak and mean pupil dilation as well as to the different terms resulting from fitting a third-order orthogonal polynomial within growth curve analysis (intercept, 1st order, 2nd order and 3rd order terms), which are assumed to provide further information about temporal changes of the pupil dilation. High values of test-retest reliability were found on some measures of the pupil response. Furthermore, a Bland-Altman analysis was applied as a graphical representation of the reliability of the pupillometry. The results showed different levels of reliability depending on the different features of the pupil response (slope, rise-fall and mean pupil dilation for the HI listeners; rise-fall, delay and mean pupil dilation for NH).

## INTRODUCTION

Pupillometry has been considered as a tool for reflecting listening effort, particularly in HI people who typically have higher listening effort than NH listeners in a given condition (Kramer *et al.*, 2006; Wendt *et al.*, 2015). Changes in listening effort as indicated by changes in the pupil size have been demonstrated on a group level (Zekveld *et al.*, 2010; Wendt *et al.*, 2015). The mentioned studies used speech-in-noise tests in combination with pupillometry to examine the impact of intelligibility, signal-to-noise ratio (SNR) and type of noise on listening effort as indicated by changes in the pupil dilation. However, the reliability of pupillometry as an indicator of individual listening effort has not been systematically studied yet.

*Corresponding author: mnea@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019.  ISSN: 2596-5522.

The current study investigated the reliability of pupillometry as an objective listening effort measure in individuals, while they perform a speech-in-noise test. The most common methods for assessing test-retest reliability are the Intraclass Correlation Coefficient (ICC), proposed by Hays *et al.* (1993), and the Bland and Altman (1986) approach. Alhanbali *et al.* (2019) showed a good reliability (ICC $>$ 0.85) of the mean and the peak pupil dilation (PPD). However, the reliability of other pupil dilation characteristics, such as time-dependent features of the pupil response was not considered in their study. Therefore, the present study focused on the reliability of the pupil dilation as a measure of listening effort by considering features such as the average height of the pupil response function, the slope, the rise and fall around the inflection point and the inflexions at the extremities of the function. These features were extracted when applying the growth curve analysis (GCA) model developed by Mirman *et al.* (2008). Furthermore, this study explored the visual representation of the reliability by using the Bland and Altman (1986) approach describing the individual differences of the two visits against their average. Another element of this study was to perform a cluster analysis on the individual responses of the pupil. The purpose of the cluster analysis was to identify the main features of the pupillary response function that could best characterize listening effort.

## METHODS

### Data set

Two different data sets were analysed as reported in Wendt *et al.* (2018) and Ohlenforst *et al.* (2018). The first data set was collected by Wendt *et al.* (2018) for a group of 27 NH listeners while the second data set was recorded by Ohlenforst *et al.* (2018) for a group of 24 HI listeners. The pupil dilation was recorded while people performed a speech-in-noise test (HINT, Nielsen and Dau (2011)) at 8 different SNRs. Only two subsets were considered for assessing reliability (two out of eight SNRs for each group, NH and HI: 0dB and 4dB, each tested at a different date) and three subsets for the cluster analysis (8dB, 0dB, -8dB for NH and HI). Four to six weeks were considered in between the two different dates, to avoid learning effects with respect to the sentence material since the sentences were repeatedly used. A list of 25 sentences per condition was presented to the participants in a block-based design. The pupil data were processed using MATLAB and R. To remove any initial effects, the first five sentences (out of 25) of the pupil traces from a list were excluded from the analysis. Data cleaning was performed as reported in Wendt *et al.* (2018). Trials with less than 80% reliable data were removed from the analysis and the other traces were baseline corrected. In total, 40 recordings of each individual were compared between the two dates (2x20x27NH, 2x20x24HI). The mean pupil dilation was calculated as the average pupil dilation over the trials. The PPD was calculated between the 3rd and 8th second of the stimulus presentation as in Zekveld *et al.* (2010).

### Growth curve analysis (GCA)

To examine temporal changes of the pupil response function for the two different dates, GCA was applied twice for the 2 different dates. According to Mirman *et al.* (2008), GCA fits orthogonal polynomial terms to time series data with the purpose of showing different variations in the function among individuals. To describe the shape of the function, three orthogonal polynomials ($p_1$, $p_2$ and $p_3$) were used. Pupil size was considered as a dependent variable in the model, predicted by a series of fixed and random effects (Eq. 1). The temporal features of the pupil response for the two dates extracted through GCA were considered when calculating test-retest reliability. According to Kalenine *et al.* (2012), the intercept term represents the averaged height of the pupil response, the linear term reflects the slope, the quadratic term reflects the rise and fall around the central inflection point of the response function, and the cubic term reflects the inflexions at the extremities of the curve referred to as delay in the current study. In other words, an estimate of the 3 coefficients and the intercept were obtained, representing the GCA terms of different orders.

$$pupil \sim (p_1 + p_2 + p_3) * participant + (1 + p_1 + p_2 + p_3 | sentence) \qquad \text{(Eq. 1)}$$

### ICC

Intraclass correlation coefficient (ICC) is one of the most used reliability indices in test-retest studies. The ICC can reflect either the degree of consistency or the agreement between measurements. The agreement assumes that the values measured on two different dates are expected to be equal for each respondent. Consistency considers that the values measured on two different occasions are correlated in an additive manner. Thus this measure is less relevant in the current analysis, but is nevertheless still reported. ICC agreement was calculated according to Hays *et al.* (1993), as reflected in Eq. 2, where $MS_R$ is the mean square for rows, $MS_C$ is the mean square for columns, $MS_E$ is the mean square for error and $n$ is the number of subjects.

$$ICC_{agreement} = \frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}} \qquad \text{(Eq. 2)}$$

### Bland-Altman (BA) approach

To apply the BA approach, the first step was to calculate the limits of agreement (LoA) as the mean $\pm$ 1.96 standard deviation of the two similarly conditioned tests. The plot is designed to show the difference between the two visits against their mean, according to Bland and Altman (1986). The bias is an important aspect in the interpretation of the BA approach, and it was calculated as a mean applied to the difference between the value determined in the first visit and the value determined in the second visit.

### Cluster analysis

The aim of applying a clustering algorithm was to identify whether the data points will group according to the different levels of SNRs, or with respect to the different

characteristics of the pupil traces from the individuals. The *k-means* (k=number of clusters) clustering algorithm applied in this study divides the data into different clusters, based on the distance between points (Euclidean distance). Given the distance between all data points and the centroids (the center of the cluster), the measurement will be assigned to the cluster with the nearest centroid.

## RESULTS

### Pupillometry data

Fig. 1 shows the pupil response of the most representative 10 (out of 27) individual NH listeners for the two test-retest pupil data sets. Significant effects were obtained on the GCA terms (intercept, linear, quadratic and cubic) with small p-values of polynomials estimates for both visits (between $1.18 \cdot 10^{-08}$- 0.009). Similarly, Fig. 2 shows the pupil response of the 10 most representative (out of 24) individual HI listeners for the two test-retest pupil data sets. Significant effects were obtained on the GCA terms as indicated by small p-values for both visits (between $5.32 \cdot 10^{-15}$- 0.012).

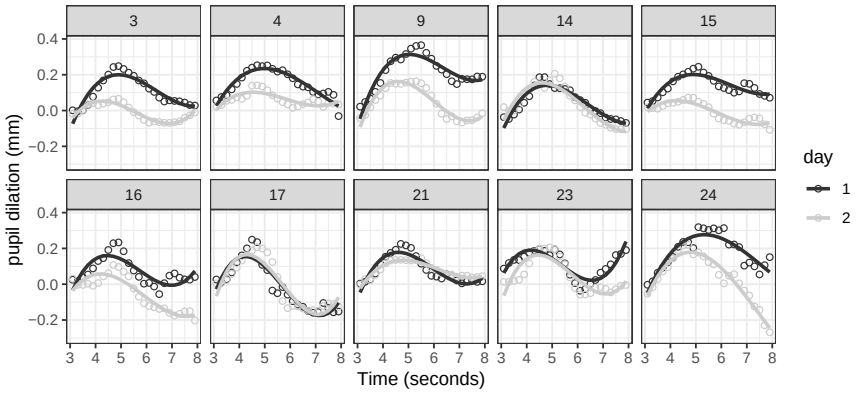



**Fig. 1:** Growth Curve Analysis for individual NH listeners. Examples of the most relevant pupil responses as a function of time, on the two different visits (black and grey). The open circles represent the actual data, while the lined functions show the fitted GCA model. The numbers in the figure represent the test subjects.

Both figures show that there were individual listeners with comparable pupil responses obtained at the two visits (e.g. NH 14, 17, 21, HI 15). However, there were also individuals showing clearly different responses (e.g. NH 9, HI 17, 19) at the two visits. The dissimilarity could be explained by the difference in the condition tested (0-4 SNR) at the two visits or by other individual factors that need to be identified.

### ICC

The classical interpretation of the ICC states that an excellent reliability is reached when ICC values are over 0.75, a good one when ICC is between 0.60 and 0.74 and a fair one for values between 0.4 and 0.59 (Chicchetti, 1994). In the current study,

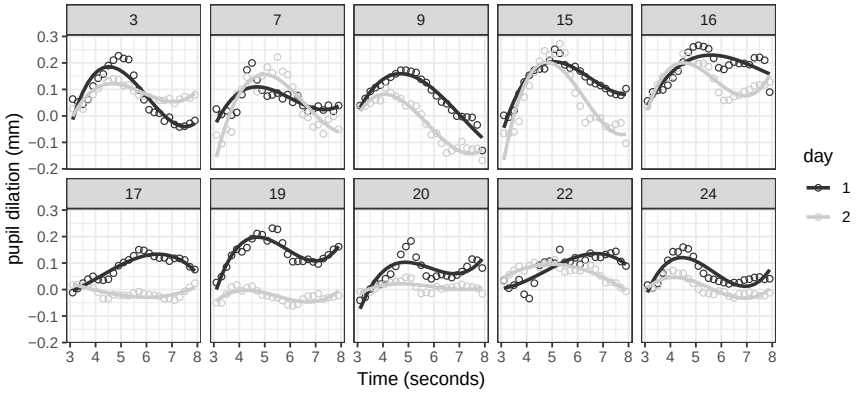



**Fig. 2:** Growth Curve Analysis for individual HI listeners. Examples of the most relevant pupil responses as a function of time, on the two different visits (black and grey). The open circles represent the actual data, while the lined functions show the fitted GCA model. The numbers in the figure represent the test subjects

the correlation coefficient was calculated for the mean, peak pupil dilation and the time-dependent terms obtained when applying the GCA model. Table 1 shows the ICC values obtained by assessing the reliability of the different features of the pupil response indicating the individual listening effort.

| ICC | NH | | HI | |
|---|---|---|---|---|
| | Agreement | Consistency | Agreement | Consistency |
| GCA Average peak | 0.6 | **0.62** | 0.41 | 0.54 |
| GCA Slope | 0.56 | 0.58 | **0.74** | **0.73** |
| GCA Rise-fall | **0.60** | **0.69** | **0.64** | **0.66** |
| GCA Delay | **0.74** | **0.86** | 0.27 | 0.47 |
| Peak pupil dilation | 0.48 | **0.60** | 0.48 | **0.64** |
| Mean pupil dilation | **0.63** | 0.59 | **0.60** | **0.64** |

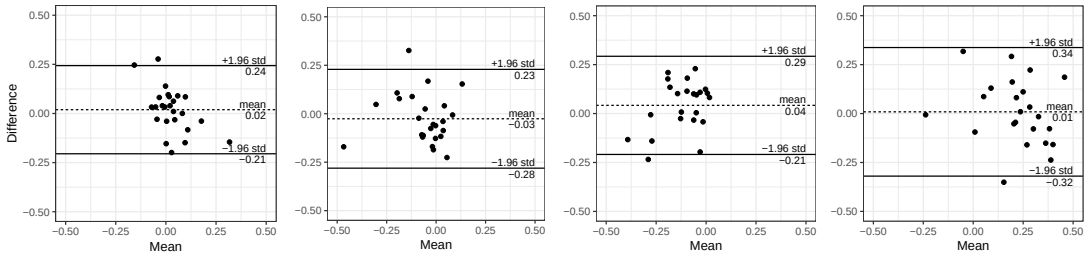
**Table 1:** ICC agreement and consistency for mean, peak pupil dilation and for different terms of GCA. The ICC values reflect test-retest reliability and bold values are the ones showing good reliability

Different features of the pupil response are reliable for the two listener groups (rise-fall, delay and mean pupil for the NH listener group; slope, rise-fall and mean pupil dilation for the HI listener group).

### Bland-Altman visual approach

Fig. 3 shows some examples of the agreement between tests taken on two separate visits as suggested by Bland-Altman. The difference between the two visits is shown against the mean of the two. Sometimes the value obtained on one visit was higher than the other, while sometimes the opposite was found. This contributes to a bias

close to zero. If it is not close to zero, the values of the two visits systematically produce different results, and this represents a low agreement of the method.

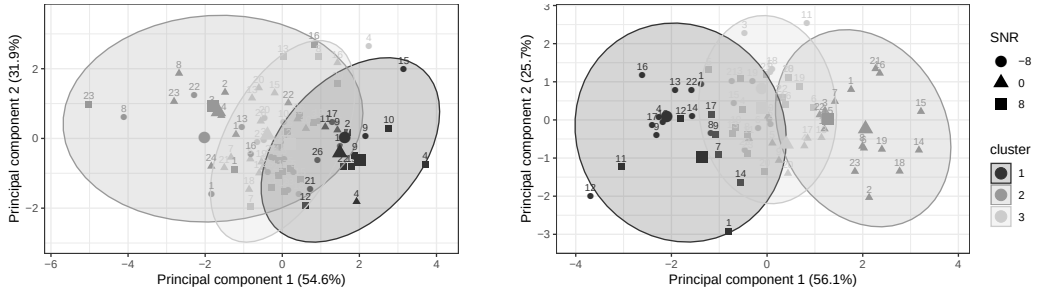



**(a)** Bland Altman Delay NH **(b)** Bland Altman Rise-fall NH **(c)** Bland Altman Rise-fall HI **(d)** Bland Altman Slope HI

**Fig. 3:** Example of Bland-Altman plots for NH (a,b) and HI (c,d) groups. The difference between two tests was plotted against their mean. 3a and 3b figures show the BA agreement for delay and rise-fall features (NH group) while the 3c and 3d figures show the BA agreement for the rise-fall and slope features (HI group).

Panels a and b of Fig. 3 show the results for the NH listeners. Most of the data points representing the delay were positioned within the LoA, as in the Fig. 3a. The bias was close to zero showing that there were no significant differences between the two visits. Panels c and d of Fig. 3 show corresponding results for the HI listeners. According to Fig. 3d, the agreement of slope was good, with large LoA values, but the bias was still close to zero. This reflects good agreement, given that the spread of the data points was broader. These results were consistent with the ICC results. Thus, the test-retest reliability was considered as good.

## Cluster analysis




**(a)** Clusters NH 4T GCA 3SNR **(b)** Clusters HI 4T GCA 3SNR

**Fig. 4:** Clustering of GCA terms for 3 different SNRs (k=3). One point represents one value of the measurement per participant per SNR.

Fig 4. shows the results of clustering the GCA terms for the NH (a) and HI (b) groups at 3 SNR conditions (-8 dB, 0 dB, 8 dB). The choice of the SNR levels to be analysed was made as in Wendt *et al.* (2018). Three different SNRs (out of the eight SNRs

contained by the dataset) with a large range between their PPD were chosen for the cluster analysis. The cluster analysis was applied to both groups, NH and HI, and the results were similar. Listeners with the same SNR were expected to be assigned to the same cluster. According to Fig. 4, the points belonging to the same cluster were data points at different SNRs, suggesting that these clusters could be formed on the base of other factors than those that were considered here.

## DISCUSSIONS AND CONCLUSION

This study showed a good reliability for some of the pupil responses features (slope, rise-fall and mean pupil dilation for the HI listeners; rise-fall, delay and mean pupil dilation for the NH listeners). The results obtained with the BA approach were consistent with the ICC results. As Alhanbali *et al.* (2019) also reported, the mean pupil size seems to be a reliable measure for both listeners groups. However, PPD was found to be less reliable than other measures in the current study. Moreover, the time-dependent features of the pupil response seem to be useful for evaluating the reliability of the method. Also, the slope seems to be more reliable for the HI group than for the NH group and it might be an important feature to explore in future studies.

The GCA model reported significant pupil features according to the small p-values of the polynomial estimates. The differences between individual functions obtained with the GCA for the two visits suggest that there could be other factors explaining the variance in the pupil curves (such as listener-dependent factors), apart from the difference in the level conditions (SNR). Zekveld *et al.* (2018) addressed some of these factors and emphasized that further investigations of the individual factors and the effects on the pupil response are required.

The cluster analysis suggested that SNR is not sufficient to classify listening effort, but that there might be some other factors needed for a classification such as listener-dependent factors like age, cognitive abilities and fatigue. Thus, future investigations of the data could consider such individual factors as input features. Furthermore, classification of the listening effort could be modeled with a supervised machine learning algorithm or even a time series analysis.

One of the limitations of the study was the use of different SNR conditions to test the pupil response reliability. It would be valuable to evaluate the reliability of pupillometry in the same acoustic conditions. Eventually, identifying and controlling the factors that can provide insights in cognitive understanding of listening situations will improve the accuracy of pupillometry as an objective measure of listening effort.

Overall, this study showed that rise-fall and mean pupil dilations seem to be important features of the pupil response, demonstrating that the signal is reliable enough in both listener groups. Other time-dependant features seemed to be reliable for one of the groups (Slope for HI and Delay for NH). The reliability results of the method are an important prerequisite for future experimental analysis and for developing pupillometry and the test protocol towards a standardized test for clinical use.

## REFERENCES

Alhanbali, S., Dawes, P., Millman, R., and Munro, K. (**2019**). "Measures of listening effort are multidimensional," Ear. Hear., **40**(5), 1084–1097, doi: 10.1097/AUD.0000000000000697.

Bland, J.M. and Altman, D.G. (**1986**). "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet., **327**(8476), 307-310, doi: 10.1016/S0140-6736(86)90837-8.

Cicchetti, D.V (**1994**). "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology.," Psychol. Assess., **6**(4), 284-290, doi: 10.1037/1040-3590.6.4.284.

Hays, R.D., Anderson, R., and Revicki, D. (**1993**). "Psychometric considerations in evaluating health-related quality of life measures," Qual. Life Res., **2**(6), 441–449, doi: 10.1007/BF00422218.

Kalenine, S., Mirman, D., Middleton, E.L., and Buxbaum, L.J. (**2012**). "Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts.," J. Exp. Psychol. Learn. Mem. Cogn., **38**(5), pp. 1274-1295, doi: 10.1037/a0027626.

Kramer, S.E., Kapteyn, T.S., and Houtgast, T. (**2006**). "Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work.," Int. J. Audiol., **45**(9), 503–512, doi: 10.1080/14992020600754583.

Mirman, D., Dixon, J.A., and Magnuson, J.S. (**2008**). " Statistical and computational models of the visual world paradigm: growth curves and individual differences," J. Mem. Lang., **59**(4), 475-494, doi: 10.1016/j.jml.2007.11.006.

Nielsen, J.B. and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol., **50**(3), 202-208, doi: 10.3109/14992027.2010.524254.

Ohlenforst, B., Wendt, D., Kramer, S.E., Naylor, G., Zekveld, A.A., and Lunner, T. (**2018**). "Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response," Hear. Res., **365**, 90-99, doi: 10.1016/j.heares.2018.05.003.

Wendt, D., Dau T., and Hjortkjær, J. (**2015**). "Impact of background noise and sentence complexity on processing demands during sentence comprehension.", Front. Psychol., **7**(31), 345, doi: 10.3389/fpsyg.2016.00345.

Wendt, D., Koelewijn, T., Książek, P., Kramer, S.E., and Lunner, T. (**2018**). "Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test," Hear. Res., **369**, 67-78, doi: 10.1016/j.heares.2018.05.006.

Zekveld, A., Kramer, S., and Festen, J. (**2010**). " Pupil response as an indication of effortful listening: the influence of sentence intelligibility.," Ear. Hear., **31**(4), 480-490, doi: 10.1097/AUD.0b013e3181d4f251.

Zekveld, A.A., Koelewijn, T., and Kramer, S.E. (**2018**). " The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge," Trends. Hear., **22**(4412), doi: 10.1177/2331216518777174.