 doi:10.20944/preprints201901.0276.v1

Article

Estimation of olfactory sensitivity using a Bayesian adaptive method

Richard Höchenberger [1,2] and Kathrin Ohla [1,2,*]

[1] Institute of Neuroscience and Medicine INM-3, Research Center Jülich, Jülich, Germany
[2] Psychophysiology of Food Perception, German Institute of Human Nutrition Potsdam-Rehbrücke, Nuthetal, Germany
* Correspondence: k.ohla@fz-juelich.de

Abstract: The ability to smell is crucial for most species as it enables the detection of environmental threats like smoke, it fosters social interactions, and it contributes to the sensory evaluation of food and eating behavior. The high prevalence for smell disturbances throughout the life span call for a continuous effort to improve tools for the quick and reliable assessment of the ability to smell. Odor-dispensing pens, called Sniffin' Sticks, are an established tool to test olfactory f unction. We tested the suitability of a Bayesian adaptive algorithm (QUEST) to estimate olfactory sensitivity using Sniffin' Sticks by comparing its results with those obtained via the established standard protocol, which relies on a staircase procedure. Thresholds were measured according to both procedures in two sessions (Test and Retest). The staircase successfully yielded threshold estimates in more cases than QUEST. Yet, Test-Retest correlations showed stronger reliability for QUEST ($\rho = 0.70$) than for staircase thresholds ($\rho = 0.50$). A strong correlation ($\rho = 0.80$) between the results of both procedures indicated good validity of QUEST. We conclude that the QUEST procedure may offer quicker convergence and reduced testing time in some cases, but fail to yield a threshold estimate in others.

Keywords: smell sensitivity; olfaction; threshold; staircase; QUEST

---

1. Introduction

The appreciation of food involves all senses: sight, smell, taste, touch, and often also hearing. While the sight of a cup of coffee may indicate its availability, it is typically its smell that is appealing and that triggers an appetite for most people. During consumption, the smell or aroma is perceived again retronasally and supported by its pleasant temperature and a bitter note. These largely parallel sensations occur automatically and only raise awareness when one or more senses are disturbed. That said, the sense of smell has been shown to influence food choice and eating behavior [1], and its impairment has even been associated with a higher risk for diet-related diseases like diabetes [2]. Given that the estimated prevalence for smell impairment is 3.5 % in the United States [3], continuous efforts are made toward an efficient and precise assessment of smell.

The *Sniffin' Sticks* test suite, developed by [4], is an established tool in the assessment of olfactory function. It consists of three tests involving sets of impregnated felt-tip pens: odor detection threshold (T), odor discrimination (D), and odor identification (I). Each test produces a number in the range from 1 to 16 as a performance measure. Overall olfactory function is assessed by summing all three test results, resulting in the *TDI score*. By comparing an individual's TDI score to the comprehensive set of available normative data (e.g. [5]), a researcher or practitioner can reliably diagnose olfactory impairment. Notably, threshold, discrimination, and identification measure different facets of olfactory function [6]. The threshold, however, has been found to explain a larger portion of variability in TDI

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2019 doi:10.20944/preprints201901.0276.v1

scores than the two other measures [7]. Moreover, the discrimination and identification tests follow relatively simple test protocols in which all stimuli are presented only once and in a pre-defined. The threshold, in comparison, is of a more complex nature, and therefore provides the largest potential for possible improvements. It follows a so-called adaptive method; specifically, a "transformed" 1-up / 2-down staircase procedure [8]. The procedure first assesses a starting concentration and then moves on to the "actual" threshold estimation, during which fixed step widths are used: for each incorrect answer, stimulus concentration is increased by one step; and for two consecutive correct answers, stimulus concentration is decreased by one step [4].

Since the 1-up / 2-down staircase was first conceived, several new approaches to threshold estimation have been published, including Bayesian methods. A Bayesian method tries to estimate parameters of the psychometric function (e.g., the threshold) using Bayesian inference: based on prior assumptions about the true position of the parameters, the next stimulus concentration is selected such that the expected information gain about the parameters is maximized. The first published Bayesian adaptive psychometric method is the QUEST procedure [9], which is still popular today. QUEST has two distinct properties that set it apart from the staircase described above. First, stimulus placement always considers the entire response history and is not solely based on the last one or two preceding trials. Second, QUEST is not tied to a fixed step width, allowing it to traverse through a large range of concentrations more quickly.

In a clinical setting, at the ENT practice or a the bedside in the hospital, shorter testing times are always beneficial, as they reduce strain on patients and free up time for other parts of diagnostics and treatment. But also when working with healthy participants, e.g. in a psychophysical lab, reduced testing time spares resources and allows for a larger number of measurements in a given time. [10] used QUEST to estimate gustatory thresholds; the method proved to converge reliably and quickly. Inspired by these results we set out to design a QUEST-based procedure for olfactory threshold estimation and to compare its performance with that of the established staircase method.

2. Materials and Methods

2.1. Participants

36 participants (32 women; median age: 29.5 years, age range: 19–61 years) completed the study. All participants were healthy and reported not having suffered from an infectious rhinitis for at least to weeks before testing. The study conformed to the revised Declaration of Helsinki and was approved by the ethical board of the German Society of Psychology (DGPs).

2.2. Stimuli

Stimuli were a set of 48 *Sniffin' Sticks*, felt-tip pens filled with an odorant (Burghart, Wedel, Germany; [4]). 16 pens were filled with different concentrations of 2-phenylethanol ranging from 4 % to approx. 1.22×10^{-4} % (a geometric sequence with the common ratio of 2, so the first pen contained a 4 % dilution, the second ½ % = 2 %; the third ⅔ % = 1 %, and so on), dissolved in 4 % propylene glycol, an odorless solvent. Note that in this test, the 1st pen contains the highest, the 16th pen the lowest odorant concentration. The remaining 32 pens were only filled with 4 % propylene glycol and served as blanks. All pens were arranged in triplets such that each triplet contained one pen with odorant and two blanks.

2.3. Procedure

2.3.1. Experimental sessions

Participants were invited for two experimental sessions – the Test and the Retest session – on different days. To ensure similar testing conditions across sessions, participants were instructed to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2019 doi:10.20944/preprints201901.0276.v1

refrain from eating, smoking, and drinking anything but water 30 min before visiting the laboratory. Further, both sessions were scheduled at approximately the same time of day, and with the shortest inter-session intervals the participants' schedules allowed for; we aimed for 7 days or less. In each session, olfactory detection thresholds were determined via the two distinct algorithms described below. Algorithm order was balanced across participants and kept constant for Test and Retest within each participant. Additionally, odor discrimination and odor identification ability were measured in one of the sessions, according to the standard *Sniffin' Sticks* protocol [4]. These data are not reported here.

2.3.2. Stimulus presentation

At the beginning of each test, participants were blindfolded. The experimenter wore odorless cotton gloves when presenting the stimuli. To present a stimulus, the experimenter removed the cap from the pen, held the tip of the pen in front of the participant's nose, approx. 2 cm from the nostrils, and asked the participant to take a sniff. Participants were informed that the odorant may be presented in very low concentrations, and that only one of the 3 pens presented in each trial contained the odorant, while the others contained the solvent exclusively. The task was to "indicate which of the three pens smells different from the others", and participants had to provide a response even when unsure. This three-alternative forced-choice task (3-AFC) yields a probability of $\frac{1}{3}$ of guessing correctly. Participants were familiarized with the odorant by presenting pen no. 1 before testing commenced. During testing, stimulus triplets were presented in intervals of approx. 20 s.

Staircase

Following the standard protocol [4], the presentation order of pens within the triplets varied from trial to trial. In the first trial, the odor pen was presented first; in the second trial, it was presented between two blanks; and in the third, after two blanks. After the third trial, this sequence was repeated.

We first determined the starting concentration. Beginning with the presentation of triplet no. 16 or 15 (balanced across participants), participants had to indicate which of the pens smelled different. Concentration was increased in steps of two (e.g., from pen 16 to 14) for each incorrect response. Once participants provided a correct response, the same triplet was presented again. If the response was incorrect, the concentration was increased again by two steps as before; however, if the triplet was correctly identified a second time, that dilution step served as the starting concentration. Contrary to the standard protocol, where testing would now continue without interruption, our participants were granted a short break of approx. 1 min before the actual threshold estimation started with the presentation of the triplet containing the starting concentration. The threshold was now determined in a 1-up / 2-down staircase procedure: odor concentration was increased by one step after each incorrect response (1-up), and decreased by one step after two consecutive correct responses at the same concentration (2-down). This kind of staircase targets a threshold of 70.71 % correct responses ([8]; but cf. [11], who found small deviations from this value). That is, if presented repeatedly with a stimulus at threshold intensity, participants would be able to correctly identify it in about 71 out of 100 cases. The probability of providing *two consecutive* correct responses purely by guessing is $\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$, assuming participants have not identified the pattern behind the pen presentation order. The procedure finished after 7 reversal points were reached. The final threshold estimate was the mean of the last 4 reversal concentrations. This procedure will be referred to simply as *staircase* throughout the rest of this manuscript.

QUEST

When using QUEST, the experimenter first has to decide upon a set of parameters that describe the assumed psychometric function linking stimulus intensity and expected response behavior. We assumed a sigmoid psychometric function of the Weibull family, as proposed by [9] and used for gustatory testing by [10], with a slope $\beta = 3.5$, a lower asymptote of $\frac{1}{3}$ (chance of a correct response

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2019 doi:10.20944/preprints201901.0276.v1

just by guessing), and a lapse rate of 0.01. This yielded a function extending from 0.34 to 0.99 in units of "proportion of correct responses". The granularity of the concentration grid was set to 0.01. All parameters of this function were constant, except for the threshold, which was the parameter of interest that was going to be estimated in the course of the procedure. The prior estimate of the threshold was a normal distribution with a standard deviation of 20, which was centered on the concentration of pen no. 7; that is, pen no. 7 was used as the starting concentration. The algorithm was set to target the threshold at 80 % correct responses, which is slightly higher than the threshold target in the staircase procedure, but had proven to produce good results both in pilot testing as well as in gustatory threshold estimation [10]. As the QUEST procedure always started with the presentation of pen no. 7, the estimation of individual starting concentrations could be omitted. Unlike in the staircase procedure, where the order of pen presentation varied in a predictable manner from triplet to triplet, here we presented the pens in random order on each trial. After a response, QUEST updates its knowledge on the expected threshold location and proposes the concentration to present in the next trial in order to maximize the expected information gain about the "true" threshold. As the set of concentrations was discrete and limited to 16, QUEST might propose concentrations we didn't have available. In that case, our software picked the triplet with the concentration closest to the one QUEST had proposed. It is noteworthy that in this procedure, the step width was not fixed as in the staircase, where the concentration was always only decreased or increased by one step. For example, QUEST might decide to step up 3 concentrations in one trial, step down 2 in the next, and present the exact same concentration again in the following trial. Whenever the same concentration had been presented on two consecutive trials, concentration was decreased if both responses had been correct, and increased if both responses had been incorrect. The procedure always ended after 20 trials. The final threshold estimate was the mean of the posterior probability density function of the threshold parameter. We will refer to this procedure as "QUEST" from now on.

2.3.3. Analysis

Data cleaning

After a participant has reached one of the "extreme" concentrations (i.e., pens no. 1 or 16) and provides a response that would, theoretically, require to present a concentration outside the prepared stimulus set, the procedure cannot be safely assumed to converge properly anymore. We therefore decided to assign a threshold value of $T = 1$ if at least one incorrect response was given for triplet no. 1, and a threshold value of $T = 16$ if one (in QUEST) or two consecutive (in the staircase) correct responses were given for triplet no. 16.

Still, the procedures may have failed to converge in some of the remaining runs, for example if response behavior was inconsistent. We therefore inspected the runs visually for obvious non-convergence and dropped the affected participants from analysis entirely (i.e., in this case, we discarded all data from both procedures, even if just a single run in only one procedure was affected). That way, way ended up with a balanced dataset, containing threshold estimates for all participants across both procedures and sessions.

Reliability

To establish reliability measures, we compared the threshold estimates of Test and Retest sessions for both procedures individually. For this, we assessed differences in session means; calculated the degree of correlation between both sessions; and fitted a linear regression model.

Since we assumed that the transformations described in the *Data cleaning* section might have introduced deviations from normality, we visually inspected Q–Q plots of the data and calculated Shapiro-Wilk test statistics. We discovered a deviation from normality for the QUEST Test thresholds ($W = 0.92$, $p < 0.05$). Therefore, comparisons of the threshold means were carried out using a Wilcoxon signed-rank test. Correlations were calculated via Spearman's rank correlation (Spearman's

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2019 doi:10.20944/preprints201901.0276.v1

rho, denoted as ρ) to estimate the degree of monotonic relationship between the measurements. Both methods are non-parametric tests that do not require the variables to be normally distributed. Ordinary least squares (OLS) models were used to fit regression lines to provide a better understanding of the nature of the relationship between the threshold estimates. Finally, to test whether the duration of the inter-session interval might be a confounding factor in the threshold estimates, we calculated the Spearman correlation between inter-session intervals and differences between Test and Retest thresholds.

Comparison between procedures

Threshold estimates fluctuate across sessions. To reduce the influence of "outlier sessions" on a participant's threshold value, we averaged Test and Retest threshold estimates for each participant within procedures. Similar to the analysis of reliability, means of those averaged thresholds were first compared using a Wilcoxon signed-rank test, followed by the calculation of Spearman's rho and the fit of a regression line using an OLS model.

Software

The experiments were run via PsychoPy 1.85.4 [12,13] running on Python 2.7.14 (https://www.python.org) installed via the Miniconda distribution (https://conda.io/miniconda.html) on Windows 7 (Microsoft Corp., Redmond, WA/USA). All analyses were carried out with Python 3.7.1, running on macOS 10.14.2 (Apple Inc., Cupertino, CA/USA). We used the following Python packages: correlation coefficients and Q-Q plots were derived via pingouin 0.2.2 [14]; Shapiro-Wilk statistics were calculated with SciPy [15,16]; linear regression models were estimated using statsmodels 0.9.0 [17]; and plots were created with seaborn 0.9.0 (https://seaborn.pydata.org) and matplotlib 3.0.2 [18].

3. Results

3.1. Data cleaning

The highest concentration, pen no. 1, was not correctly identified in 5 runs (5 participants) during the staircase, and 12 times (11 participants) during the QUEST procedure. Accordingly, these thresholds were assumed to be $T = 1$. None of the participants ever provided correct responses at the lowest concentration, pen no. 16. Visual inspection indicated that two QUEST runs had not properly converged (2 participants: both women, aged 26 and 28 years), and these participants were thus excluded from all analysis, leaving a total of 34 participants that entered analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2019 doi:10.20944/preprints201901.0276.v1

3.2. Reliability

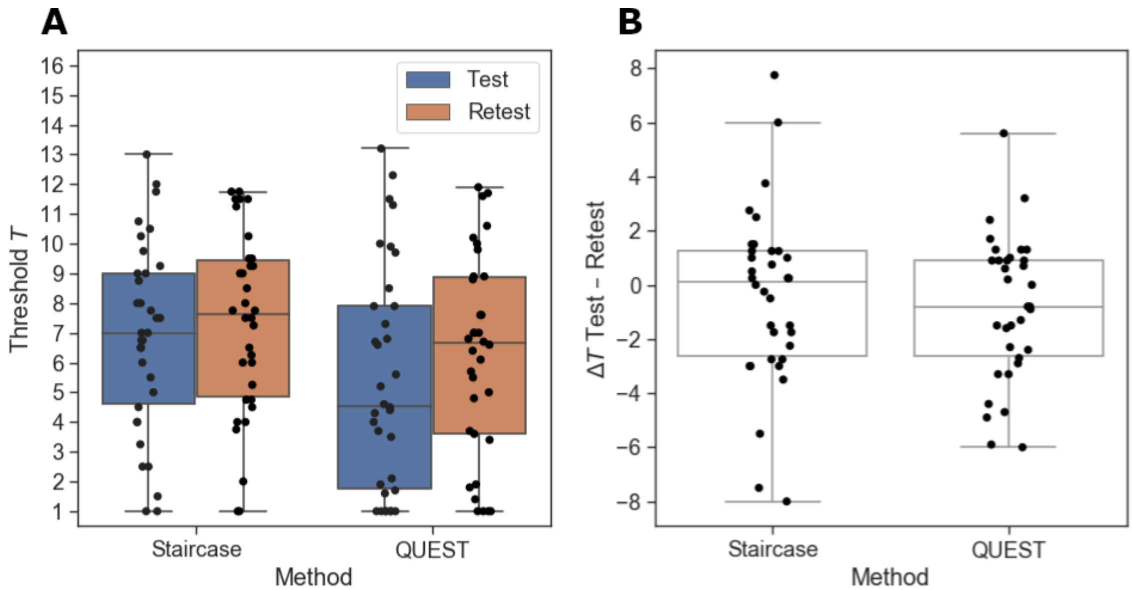

Figure 1. (**A**) Threshold estimates for the staircase and QUEST procedures during Test and Retest sessions. (**B**) Differences between Test and Retest threshold estimates. Each data point represents one participant. Whisker length represents 1.5 × inter-quartile range.

Mean Test and Retest thresholds did not differ for the staircase ($M_{\text{Test}} = 6.9$, $\text{SD}_{\text{Test}} = 3.2$; $M_{\text{Retest}} = 7.3$, $\text{SD}_{\text{Retest}} = 3.1$; $W = 231.0$, $p = 0.14$), but there was a significant difference for QUEST ($M_{\text{Test}} = 5.4$, $\text{SD}_{\text{Test}} = 3.8$; $M_{\text{Retest}} = 6.2$, $\text{SD}_{\text{Retest}} = 3.4$; $W = 187.5$, $p < 0.01$; see Fig. 1 A). The *differences* between Test and Retest thresholds were more dispersed for the staircase than for QUEST ($\text{SD}_{\Delta,\text{staircase}} = 3.26$; $\text{SD}_{\Delta,\text{QUEST}} = 2.65$; see Fig. 1 B).

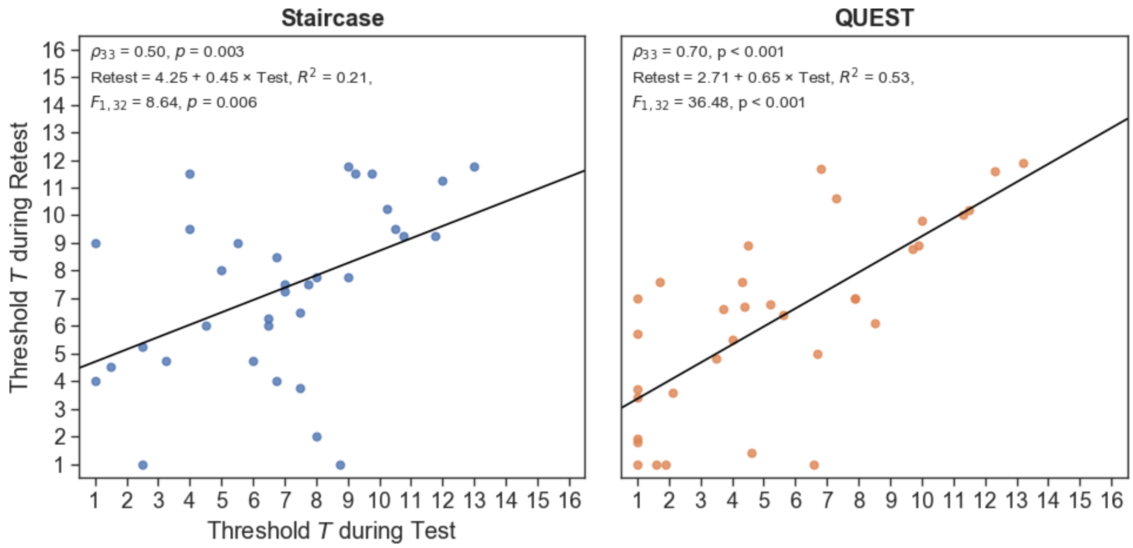

Figure 2. Correlation between Test and Retest threshold estimates.

The thresholds estimated for Test and Retest sessions correlated significantly for both procedures, with QUEST demonstrating a stronger correlation than the staircase (staircase: $\rho_{33} = 0.50$, $p < 0.01$; QUEST: $\rho_{33} = 0.70$, $p < 0.001$; see Fig. 2).

Considering that during the data cleaning procedure 12 QUEST, but only 5 of the staircase thresholds had been assumed to be "1" as participants had provided incorrect responses at the highest

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2019 doi:10.20944/preprints201901.0276.v1

concentration, we re-ran the analysis, but with all participants removed who had failed to identify pen no. 1. Now correlation coefficients were much more similar between both procedures, and slightly higher for the staircase compared to QUEST (staircase: $\rho_{22} = 0.73$, $p < 0.001$; QUEST: $\rho_{22} = 0.68$, $p < 0.001$). This approach, however, may have excluded particularly "difficult" participants, so this result should be considered with caution; it was merely calculated for exploratory purposes and to provide the reader with a better idea of the influences our data cleaning procedure had on some of the results presented here.

Inter-session intervals were relatively short (median: 3.0 days; range: 0.9–8.9 days). Only two participants exceeded the intended 7-day interval limit (8.0 and 9.0 days, respectively). The difference between Test and Retest threshold estimates did not correlate with the time between sessions (staircase: $\rho_{33} = -0.06$, $p = 0.76$; QUEST: $\rho_{33} = 0.03$, $p = 0.86$).

3.3. Comparison between procedures

The mean threshold estimates (i.e., averaged across sessions) for the staircase were higher and varied less than for QUEST (staircase: $M = 7.1$, SD = 2.7; QUEST: $M = 5.8$, SD = 3.3; Fig. 3 A). This difference was highly significant ($W = 101.0$, $p < 0.001$). Yet, the thresholds correlated significantly ($\rho_{33} = 0.80$, $p < 0.001$), and the regression slope was very close to 1, indicating a good agreement across procedures (Fig. 3 B).

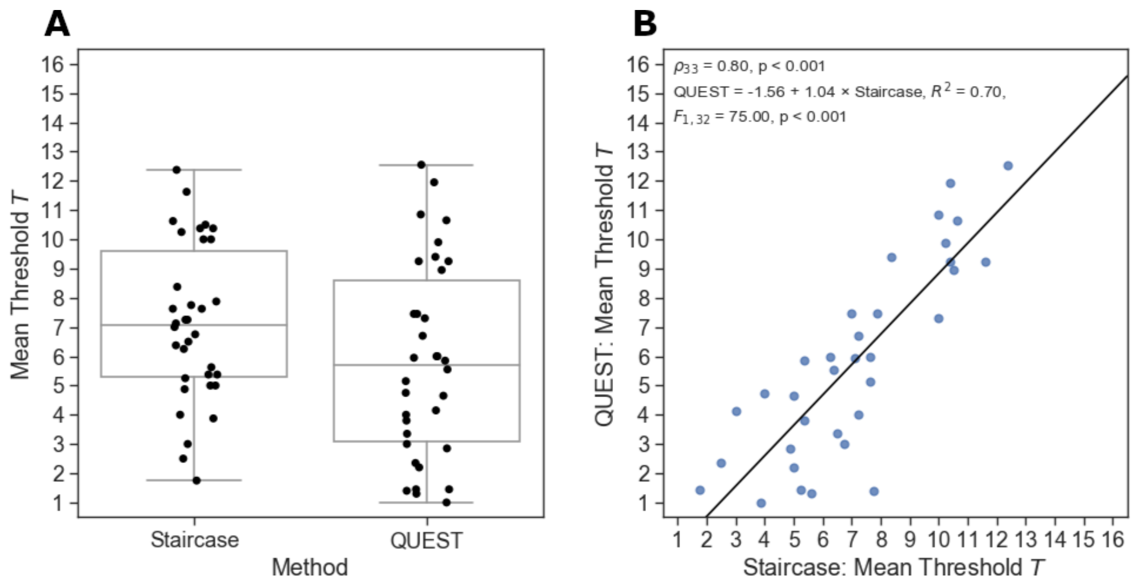

Figure 3. Comparison between thresholds estimated using the staircase and the QUEST procedure. (**A**) Mean threshold estimates, averaged across sessions. (**B**) Correlation between mean staircase and QUEST threshold estimates. Each data point represents one participant. Whisker length represents 1.5 × inter-quartile range.

4. Discussion

In the presented study we used a QUEST-based algorithm to estimate olfactory detection thresholds for 2-phenylethanol. The aim was to provide a reliable test result as it had recently been demonstrated for taste thresholds [10] and, ideally, with reduced testing time. The results were compared to the widely-used testing protocol based on a 1-up / 2-down staircase procedure [4–6,19].

We found good test-retest reliability the QUEST procedure ($\rho = 0.70$). In contrast, reliability of the staircase procedure was only moderate ($\rho = 0.50$) and lower than reported in previous studies for n-butanol ($r = 0.61$ [4]) and 2-phenylethanol ($r = 0.92$ [6]) thresholds. These studies however, tested larger samples with a more balanced gender distribution, while almost 90 % of our participants were women. Although neither a previous study with several hundred participants [19], nor a more recent investigation involving more than 3,000 participants [5] could find any gender effects in n-butanol

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2019 doi:10.20944/preprints201901.0276.v1

thresholds assessed via the standard *Sniffin' Sticks* procedure, it cannot be excluded that a gender bias contributed, at least partially, to our results. [4] reported better performance (significantly lower thresholds, i.e., higher pen numbers) in the second session, compared to the first. The results from the QUEST procedure align well with this observation; the staircase, too, yielded lower thresholds in the second session, albeit the difference was not significant.

Comparison of the mean thresholds (averaged across the two sessions) revealed a strong correlation between the procedures, and regression analysis showed an almost perfect linear relationship, demonstrating a good agreement between QUEST and staircase results. Notably, the staircase yielded slightly higher pen numbers (i.e., lower thresholds) than QUEST. This was expected as the procedures were assumed to converge at approx. 71 % and 80 % correct responses, respectively.

Surprisingly, a number of participants were unable to correctly identify pen no. 1, and this effect was more pronounced during QUEST compared to the staircase. Theoretically, the variable step sizes used by QUEST render it possible to quickly approach even the extreme concentration ranges. Visual inspection of the trial and response sequences of QUEST runs in which participants failed to identify pen no. 1, however, provided no clear indication that the variability in step sizes led to an implausible sequence of stimulus presentations. Because some participants provided both, correct and incorrect responses when presented with pen no. 1 repeatedly, the current criterion of assigning a threshold $T = 1$ after a single failure to identify the pen might be too strict. Loosening of this criterion could, however, lead to threshold estimates of "virtual" pens below 1 in some cases, so it is questionable whether this approach would produce additional information of value.

QUEST successfully converged within 20 trials for most participants. This gives QUEST an advantage in some situations, where threshold estimation may finish quicker than with the staircase procedure. The QUEST procedure could be further optimized by introducing a dynamic stopping rule. For example, [10] set the algorithm to terminate once the threshold estimate had reached a certain degree of confidence. Such a rule can further reduce testing time, as the run may finish in fewer than 20 trials, and should be considered in future studies.

During analysis of the data we discovered that a number of the staircase runs seemed to have not fully converged although 7 reversal points were reached. This commonly happened in runs where participants seemed to exhibit a somewhat "fluctuating" threshold that caused the procedure to move in the direction of higher concentrations throughout the procedure, but without ever reaching pen no. 1. In some of these cases, QUEST proved to behave more consistently by either converging to a threshold or actually reaching pen no. 1, which would then sometimes not be identified correctly. These interesting differences in behavior require further investigation to fully understand their cause and influence on threshold estimates and, ultimately, diagnostics.

5. Conclusions

We applied a procedure based on the QUEST algorithm to estimate olfactory detection sensitivity. The algorithm proved to produce reliable results which differed systematically, but reliably, from those acquired using an established staircase protocol. Overall, the measurement results of both procedures largely overlapped. The QUEST algorithm may offer reduced testing time and better convergence in some cases, but fail to yield an actual threshold estimate in others. Further research is needed to better understand possible advantages and caveats of the QUEST procedure compared to the staircase testing protocol.

6. Data and software availability

The data analyzed in this paper is available from https://doi.org/10.5281/zenodo.2548621. The authors provide a hosted service for running the presented experiments online at https://sensory-testing.org; the sources of this online implementation can be retrieved from https://github.com/hoechenberger/webtaste.

 doi:10.20944/preprints201901.0276.v1

Author Contributions: conceptualization, R.H. and K.O.; programming, analysis, and visualization, R.H.; interpretation and writing, R.H. and K.O.; supervision and project administration, K.O.

Funding: The implementation of the online interface was supported by Wikimedia Deutschland, Stifterverband, and Volkswagen Foundation through an Open Science Fellowship granted to R.H.

Acknowledgments: The data was acquired at the German Institute of Human Nutrition. The authors would like to thank Andrea Katschak for data collection.

Conflicts of Interest: The authors declare no conflict of interest. The funding agents had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

1. Boesveldt, S.; Bobowski, N.; McCrickerd, K.; Maître, I.; Sulmont-Rossé, C.; Forde, C.G. The changing role of the senses in food choice and food intake across the lifespan. *Food Quality and Preference* **2018**, *68*, 80–89. doi:10.1016/j.foodqual.2018.02.004.
2. Rasmussen, V.F.; Vestergaard, E.T.; Hejlesen, O.; Andersson, C.U.N.; Cichosz, S.L. Prevalence of taste and smell impairment in adults with diabetes: A cross-sectional analysis of data from the National Health and Nutrition Examination Survey (NHANES). *Primary Care Diabetes* **2018**, *12*, 453–459. doi:10.1016/j.pcd.2018.05.006.
3. Liu, G.; Zong, G.; Doty, R.L.; Sun, Q. Prevalence and risk factors of taste and smell impairment in a nationwide representative sample of the US population: a cross-sectional study. *BMJ Open* **2016**, *6*, e013246. doi:10.1136/bmjopen-2016-013246.
4. Hummel, T.; Sekinger, B.; Wolf, S.; Pauli, E.; Kobal, G. 'Sniffin' Sticks': Olfactory Performance Assessed by the Combined Testing of Odour Identification, Odor Discrimination and Olfactory Threshold. *Chemical Senses* **1997**, *22*, 39–52. doi:10.1093/chemse/22.1.39.
5. Hummel, T.; Kobal, G.; Gudziol, H.; Mackay-Sim, A. Normative data for the "Sniffin' Sticks" including tests of odor identification, odor discrimination, and olfactory thresholds: an upgrade based on a group of more than 3,000 subjects. *European Archives of Oto-Rhino-Laryngology* **2007**, *264*, 237–243. doi:10.1007/s00405-006-0173-0.
6. Haehner, A.; Mayer, A.M.; Landis, B.N.; Pournaras, I.; Lill, K.; Gudziol, V.; Hummel, T. High Test-Retest Reliability of the Extended Version of the "Sniffin' Sticks" Test. *Chemical Senses* **2009**, *34*, 705–711. doi:10.1093/chemse/bjp057.
7. Lötsch, J.; Reichmann, H.; Hummel, T. Different Odor Tests Contribute Differently to the Evaluation of Olfactory Loss. *Chemical Senses* **2008**, *33*, 17–21. doi:10.1093/chemse/bjm058.
8. Wetherill, G.B.; Levitt, H. Sequential Estimation of Points on a Psychometric Function. *British Journal of Mathematical and Statistical Psychology* **1965**, *18*, 1–10. doi:10.1111/j.2044-8317.1965.tb00689.x.
9. Watson, A.B.; Pelli, D.G. Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics* **1983**, *33*, 113–120. doi:10.3758/bf03202828.
10. Höchenberger, R.; Ohla, K. Rapid Estimation of Gustatory Sensitivity Thresholds with SIAM and QUEST. *Frontiers in Psychology* **2017**, *8*. doi:10.3389/fpsyg.2017.00981.
11. García-Pérez, M.A. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research* **1998**, *38*, 1861–1881. doi:10.1016/s0042-6989(97)00340-4.
12. Peirce, J.W. PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods* **2007**, *162*, 8–13. doi:10.1016/j.jneumeth.2006.11.017.
13. Peirce, J.W. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics* **2008**, *2*. doi:10.3389/neuro.11.010.2008.
14. Vallat, R. Pingouin: statistics in Python. *Journal of Open Source Software* **2018**, *3*, 1026. doi:10.21105/joss.01026.
15. Oliphant, T.E. Python for Scientific Computing. *Computing in Science & Engineering* **2007**, *9*, 10–20. doi:10.1109/mcse.2007.58.
16. Millman, K.J.; Aivazis, M. Python for Scientists and Engineers. *Computing in Science & Engineering* **2011**, *13*, 9–12. doi:10.1109/mcse.2011.36.

 doi:10.20944/preprints201901.0276.v1

17. Seabold, S.; Perktold, J. Statsmodels: Econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference. SciPy society Austin, 2010, Vol. 57, p. 61.
18. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **2007**, *9*, 90–95. doi:10.1109/mcse.2007.55.
19. Kobal, G.; Klimek, L.; Wolfensberger, M.; Gudziol, H.; Temmel, A.; Owen, C.M.; Seeber, H.; Pauli, E.; Hummel, T. Multicenter investigation of 1,036 subjects using a standardized method for the assessment of olfactory function combining tests of odor identification, odor discrimination, and olfactory thresholds. *European Archives of Oto-Rhino-Laryngology* **2000**, *257*, 205–211. doi:10.1007/s004050050223.