Active Learning for Visual Acuity Testing

Luis A. Lesmes[*]
Adaptive Sensory Technology, Inc.
San Diego, CA
luis.lesmes@adaptivesensorytech.com

Michael Dorr[†]
Adaptive Sensory Technology, Inc.
San Diego, CA
michael.dorr@adaptivesensorytech.com

ABSTRACT

We present Quantitative Visual Acuity (qVA), a novel active learning algorithm to assess visual acuity. It uses Monte Carlo simulations and an information maximization strategy during stimulus selection, and Bayesian inference to iteratively update the best estimate of the true underlying function. Compared to the state of the art, qVA uses a richer model for observer behaviour, and we use simulations to show its excellent test-retest repeatability and ability to detect change. In simulations of clinical studies with 50 "control" subjects demonstrating no visual change, and 50 "treatment" subjects demonstrating a 0.10 logMAR change (corresponding to one line of the gold-standard ETDRS letter chart), the qVA detected visual change with an AUC of 93%, relative to 78% performance by the ETDRS standard, given the same number of presented letters.

CCS CONCEPTS

• **Applied computing → Consumer health**; **Health informatics**;

KEYWORDS

Bayesian inference, active learning, health care, visual function assessment

ACM Reference Format:
Luis A. Lesmes and Michael Dorr. 2019. Active Learning for Visual Acuity Testing. In *Proceedings of Applications of Intelligent Systems (APPIS'19)*. ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Recent breakthrough progress in the development of intelligent systems is promising to radically alter many application domains, including healthcare. Data-driven approaches such as deep learning may realize the potential of personalized medicine by extracting meaningful patterns from both structured and unstructured data sources such as gene and protein expression data [2] on the one hand and clinical records [11] and behavioural data coming from fitness trackers [12] or social media platforms [5] on the other hand.

Another important potential source of rich clinical information is imaging data [18]. In ophthalmology, for example, optical coherence tomography (OCT) provides a mapping of the retinal layers and their thickness, and may be used to detect and localize lesions or abnormal blood vessel growth, which is a symptom of diseases that are among the leading causes of blindness, e.g. age-related macular degeneration and diabetic retinopathy. Machine Learning systems have recently begun to outperform human raters in the interpretation of OCT images [6].

However, even though the detection and assessment of structual changes in retinal tissue is of obvious interest to researchers and clinicians, the most critical question cannot – as of yet – be answered by structural imaging alone: How well does the patient see?

To answer this question, the current standard for behavioural visual function assessment is visual acuity (VA) testing (see Fig. 1). VA describes the smallest (full-contrast) stimulus size at which the stimulus (typically, a letter) can still be recognized; because of the probabilistic nature of the *psychometric function* (see Fig. 2), "recognition" is defined by a threshold (e.g. 60%) on the probability of a correct response. This probabilistic threshold also is the reason why traditional paper-based letter charts, where a patient has to read down a set of lines of progressively smaller letters, are necessarily imprecise: In order to obtain a robust estimate of the threshold, many letters have to be presented near the threshold, but on paper charts the majority of read letters is sized well above threshold.

While computerized adaptive vision tests exist that place more letters near threshold (specifically, the continuously updated estimate of the threshold, based on test history), e.g. [1], these present single letters instead of chart rows, use a simplified model for the psychometric function with a fixed range, and are not commonly used in clinical practice or clinical trials, where precise assessment of visual function is paramount to evaluate the natural progression of disease and the efficacy of ophthalmic interventions.

In this paper, we present a new algorithm, Quantitative Visual Acuity, that uses an Active Learning procedure to simultaneously estimate the threshold and range parameters of the psychometric function (see Fig. 2). Based on simulations and in comparison to the current gold standard of clinical trial visual function assessment, we show that qVA rapidly converges to an accurate estimate of the true underlying function, and that the qVA output is highly reliable across runs, for both standard and novel measures of repeatability. Finally, we use ROC analysis to show that qVA provides excellent specificity and sensitivity to small changes in (simulated) visual function, demonstrating its utility for clinical trials and clinical practice.

[*] These authors contributed equally.
[†] These authors contributed equally.

APPIS'19, January 2019, Las Palmas de Gran Canaria
© 2019 Copyright held by the owner/author(s).
ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

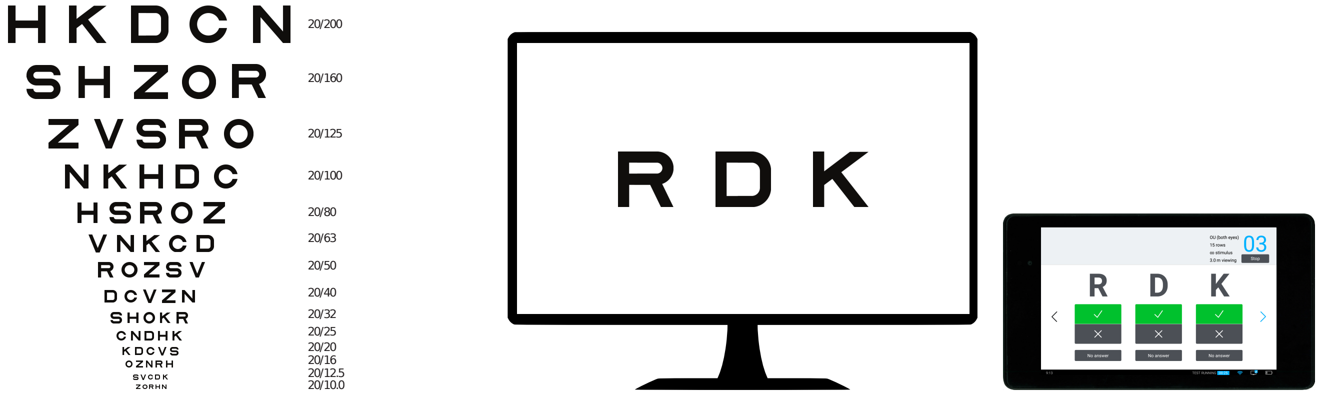

Figure 1: ETDRS eye chart (left) and a schematic of Quantitative Visual Acuity (right). The ETDRS chart is read from top to bottom until the first line with at least three mistakes, which means that test time (for most subjects) is wasted in the trivially easy upper part, but few letters are presented near the threshold of legibility. In qVA, only one three-letter row ('trial') is presented at a time on a computer monitor, and a new line (with optimal letter size for this particular subject, given the test history) is presented after the subject's responses have been entered on a tablet device.

2 METHODS

2.1 Early Treatment Diabetic Retinopathy Study Chart

The state of the art of behavioural visual function assessment in clinical trials is still the Early Treatment Diabetic Retinopathy Study (ETDRS) chart [9]. Available in different versions to reduce the risk of memorization across runs, it shows (from top to bottom) progressively smaller rows of five Sloan letters each. Because of the logarithmic relationship between stimulus intensity and perceptual experience [15], the letter size per row decrements logarithmically by 26%, or 0.1 logMAR, in each step, beginning with a letter size of 1.0 logMAR at the top.

The zero line on this chart corresponds to a minimum angle of resolution of 1 arcmin. While this is often regarded as "normal" vision (also expressed by the alternative formulation of "20/20 vision", being able to see at a distance of 20 feet what the reference observer can discern at the same distance; 20/20=1.0, and $\log_{10}1.0=0.0$ logMAR), the majority of subjects (possibly with optical correction, e.g. habitually worn glasses or contact lenses) achieve VA scores better than 0.0 logMAR; excellent vision corresponds to VA scores of $\leq$-0.3 logMAR [8, 16].

During ETDRS testing, the subject begins to read the chart from the top and continues until he or she encounters the first row where at least three (out of five) letters cannot be recognized anymore. For each correctly recognized letter on the chart, a score of 0.02 (five letters per row correspond to 0.1 logMAR) is subtracted from the baseline score of 70 to arrive at the overall score.

2.2 Quantitative Visual Acuity

2.2.1 Model of the Psychometric Function. Psychometric functions that map sensory stimuli to human behavioural responses are characterized by i) a region where the stimulus signal is too weak (e.g. the presented letters are too small) for the observer to recognize the stimulus, so that performance is at chance level; ii) a region where the stimulus signal is strong enough for the observer to reliably recognize the stimulus every time (performance at 100%); and iii) an intermediate, transition region that is typically assumed to have a sigmoidal shape and in which the observer is uncertain about the stimulus identity.

The psychometric function at the core of qVA is visualized on the left side of Fig. 2. It is parameterized by Θ, the threshold at which recognition performance is at 67%, and the range ρ, which indicates the width of the psychometric function from its 33% to its 98.2% points. A larger range indicates a shallower slope of the psychometric function, i.e. a larger region of uncertainty. Formally, the psychometric function Ψ' with threshold Θ and range ρ is defined as returning the signal detection theory sensitivity index d' for a stimulus of size τ,

$$\log_{10}\Psi'(\tau,\Theta,\rho) = \log_{10}(6) + \frac{\delta}{2\rho}(\tau-\Theta) - \frac{1}{2}\log_{10}\left(8 + 10^{\frac{\delta}{\rho}(\tau-\Theta)}\right)$$

with the constant $\delta = \log_{10}(35) - \log_{10}(1.25)$. The d' can be further converted to Ψ_n, i.e. the probabilities of different responses (zero – three correct out of three letters per row), see the right side of Fig. 2. These probabilities are used to adjust stimulus size for each row presentation during the test as described in the following subsection.

2.2.2 Active Learning. The procedure starts at trial $t = 0$ with a uniform prior $P_0(v)$ over a grid of parameter sets $v_i = (\Theta_i, \rho_i)$; in the current implementation, $\Theta \in [-0.5, 2.0]$ with 1001 equidistant steps, and $\rho \in [0.1, 1.2]$ with 51 log-equidistant steps. In each trial, qVA picks the most informative out of 100 possible stimulus sizes (ranging from -0.4 to 1.6 logMAR in 0.02 logMAR steps) and presents three randomly sampled Sloan letters of that size to the observer.

With the response r_t that encodes the number of correctly recognized letters of size τ (where the probability of r is determined by the psychometric function Ψ) and the set of previous responses $r_{0,\ldots,t-1}$, the belief over $P(v)$ is iteratively updated according to

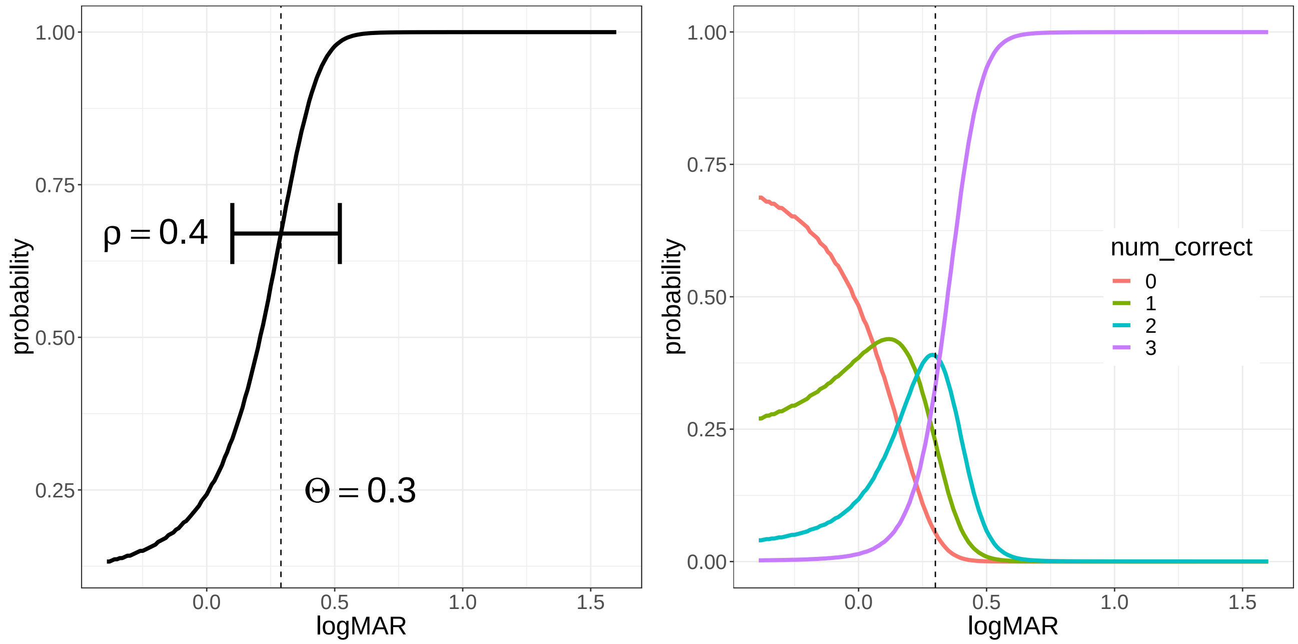

Figure 2: Single-letter psychometric function with threshold $\Theta = 0.3$ and range $\rho = 0.4$ (left) and the corresponding probability functions for different numbers of correct responses when three letters are presented (right). For a single letter, the probability of correct identification is at chance level (here, with 10 letters to choose from, 10%) for small letter sizes, i.e. $\ll 0$ logMAR, and at ceiling (100%) for very large letter sizes, i.e. > 0.5 logMAR. At threshold size (here, 0.3 logMAR), the probability of a correct response is 67%. For a row of three letters, the most likely number of correctly identified letters at threshold size is thus two, but all other responses are also possible.

Bayes' Rule,

$$\begin{aligned} p_t(v_i | r_t, r_{0,\ldots,t-1}, \tau) \\ &= p_t(r_t, r_{0,\ldots,t-1}, \tau | v_i) \cdot \frac{p_{t-1}(v_i)}{p_t(r_t, r_{0,\ldots,t-1}, \tau)} \\ &= p_{t-1}(v_i | r_t, r_{0,\ldots,t-1}, \tau) \cdot \frac{p(r_t, \tau | v_i)}{\sum_j p_{t-1}(v_i) p(r_t, \tau | v_j)}. \end{aligned}$$

Critically, the most informative stimulus size is determined by a Monte Carlo simulation of likely outcomes [13], based on a sample of 1000 grid nodes v (randomly drawn from the posterior of the previous time step, $P_{t-1}(v)$), and the expected information gain for each stimulus size τ, i.e. the reduction in entropy H of $P(v)$:

$$I_t(r_t(\tau); v) = H\left(\int_v p_t(v) \cdot \Psi_v(\tau) dv\right) - \int_v p_t(v) \cdot H(\Psi_v(\tau)) dv.$$

2.3 Simulations

We simulated 100 observers whose acuity parameters Θ and ρ were randomly sampled from the empirical posterior distributions of qVA assessments (with 45 rows) of a cohort of young adults of ocular health who were tested under corrected and blurred vision conditions. This sample gave us a range from excellent to moderate visual performance (Θ in [-0.24, 0.43] logMAR, $\mu = 0.0$, ρ in [0.09, 0.53] logMAR, $\mu = 0.26$), and empirically plausible combinations of Θ and ρ.

Because the ETDRS chart is designed to estimate a threshold at a slightly different probability correct rate (three out of five letters, i.e. 60%, vs. the qVA's two out of three, i.e. 67%), the threshold parameter Θ of the simulated observers was slightly adjusted upwards for the ETDRS simulations (depending on ρ, up to $\approx$0.03 logMAR).

To assess repeatability, we simulated five "baseline" runs for each method. For qVA, 100 trials (of three letters each) were simulated per run; for ETDRS, runs were simulated until the termination criterion (first row with less than three letters correct) held, which on average occured after 11.8 rows (58.9 letters).

To assess the sensitivity to change, each observer was simulated for another six "change" runs with an upward change in the threshold parameter Θ (i.e. worsened vision) by 0.01, 0.02, 0.03, 0.05, 0.07, and 0.1 logMAR. Based on correlations in the empirical data, the range parameter ρ was also changed, specifically by one-third the amount of the change in Θ.

2.4 Performance Analysis

2.4.1 Repeatability and Precision. According to ISO norm 5725-1, the "precision" of a measurement device (such as a vision test) is the similarity of repeated measurements of the same signal (observer) to each other, and we therefore (and in line with the clinically well-established Bland-Altman analysis [3]) calculated the standard deviation of the test-retest differences. In order to make the resulting numbers comparable to those in the literature, we computed the standard deviation for all possible pairings of the five runs into test and retest, and report the mean over these standard deviations.

However, a comparison of different tests by the test-retest standard deviation is vulnerable to artefacts because it assumes homoscedasticity, is sensitive to scaling of the test score, and may be reduced by quantization and floor and ceiling effects.

We therefore also follow the concept of "precision" as it is used in Information Retrieval. Intuitively, we want to obtain the same test score for repeated tests of the same subject, but different test scores for different subjects; in other words, a test score should be able to identify a subject within a larger population. In Information Retrieval terms, we use a subject's test score as the query and try to retrieve this subject's retest score from among all retests by their similarity to the query; the Mean Average Precision (MAP) [7] then is the precision (the share of relevant retrieved items, i.e. the inverse of the rank of this subject's retest score) for which recall is 1 (the same subject's retest has been found). We repeat this procedure for each pair of test runs and for each subject and report the grand mean.

A further advantage of MAP is that it is straightforward to extend it to multidimensional test signals. Here, we compute the similarity needed for ranking by calculating the Euclidean distance between tests in the twodimensional space of the parameters Θ and ρ. Because these parameters have different numerical ranges and different reliability, we linearly scaled them with weights that were found by crossvalidation using the data from the first two runs; the mean MAP for test-retest combinations of the remaining three runs is reported.

2.4.2 Change Detection Performance. Reliability may be a desired test property if exact absolute test scores are of interest. However, for most real-world applications, much more important than absolute scores is the assessment of relative scores, i.e. the sensitivity of a test to a change in the true underlying visual function, for example the change due to fitting different contact lenses or the progression or remediation of disease over time or due to an intervention [10].

We therefore performed Receiver-Operator Characteristic analysis for small changes in Θ and ρ and calculated the area under the ROC curve for a summary statistic of both sensitivity and specificity. In line with clinical trial design and to have a yardstick for random test-retest noise effects, we split our simulated observers into two groups.

The first 50 subjects of our simulated cohort were assigned to the "control" group and the vision "changes" were computed for the first vs. second baseline run (corresponding to test-retest repeatability above). In contrast, the second half were assigned to the "intervention" group and score changes were computed between the first baseline run and the corresponding change run.

3 RESULTS

Results of the repeatability analysis can be seen in Fig. 3 and Fig. 4. The baseline performance of the threshold estimate of the ETDRS letter chart (dashed lines) was surpassed after three and seven trials for MAP and s.d., respectively, and both repeatability measures of the qVA continued to improve (s.d. decreased, MAP increased) with an increasing number of trials. The range parameter ρ is much

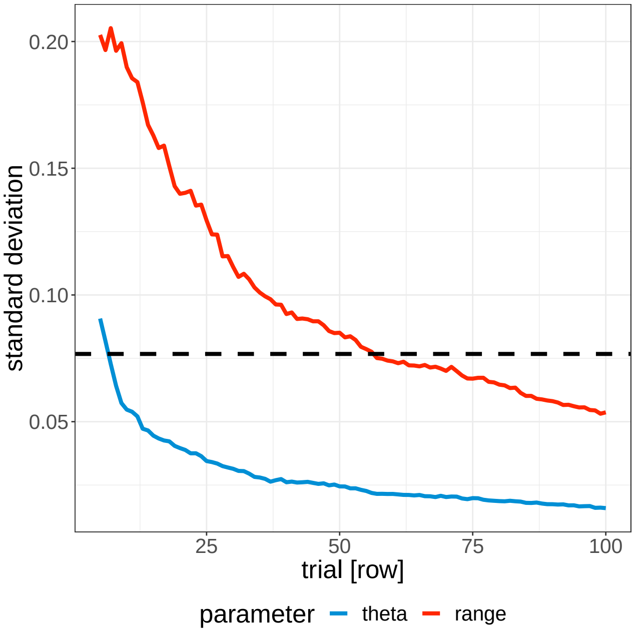

Figure 3: Repeatability (s.d. of test-retest differences) of qVA plotted as a function of number of trials. Dashed line indicates ETDRS chart baseline performance at the average number of ETDRS letters (58.9).

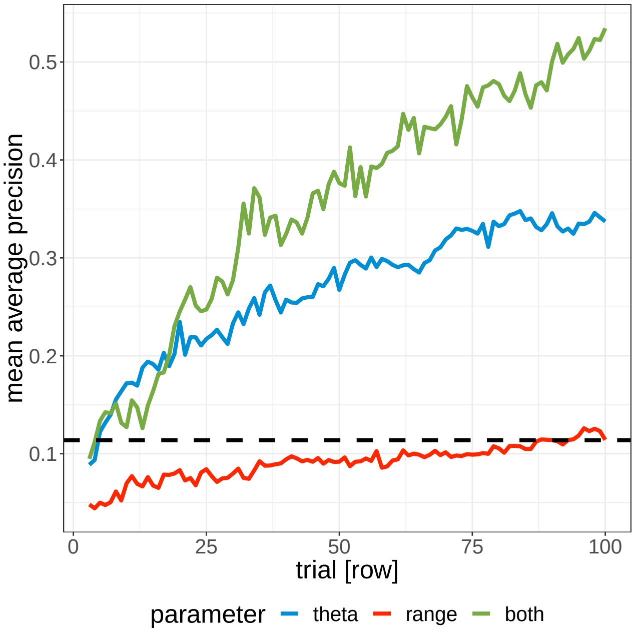

Figure 4: Mean Average Precision of qVA plotted as a function of number of trials. Dashed line indicates ETDRS chart baseline performance at the average number of ETDRS letters (58.9).

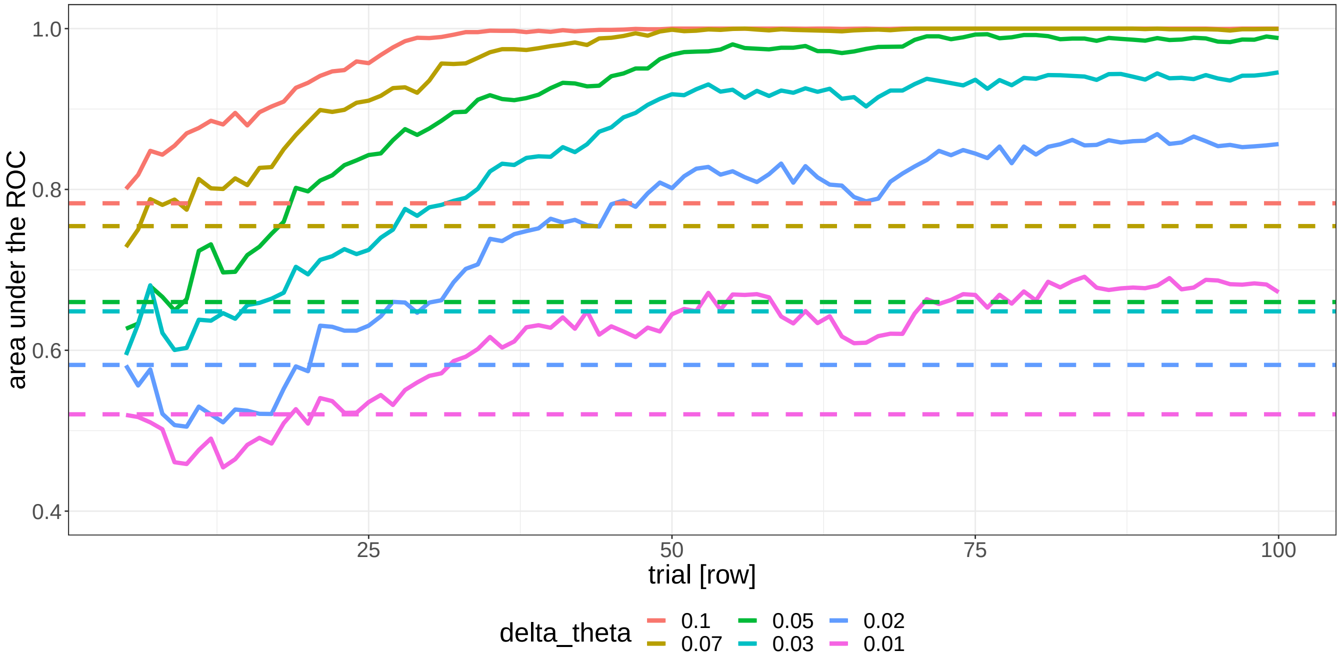

Figure 5: Area under the ROC curve results for detection of changes in visual function by qVA as a function of number of trials. Dashed lines indicate performance for the ETDRS chart at the average number of ETDRS trials.

harder to constrain and required a larger number of trials. Notably, its test-retest standard deviation reached the level of ETDRS's threshold estimation after about 55 trials, but it took about 85 trials to reach the same MAP performance.

Because of the (relative) unreliability of the range parameter ρ, the linear combination of Θ and ρ ("both" in Fig. 4) at first reduces MAP. However, after about 20 trials, the multidimensional analysis provided a benefit that continued to increase with the number of trials.

The results for change detection performance are shown in Fig. 5. The dashed lines indicate ETDRS performance (after 58.9 letters on average) and it can be seen that with ETDRS, even the largest change (of 0.1 logMAR) can be detected with an AUC of only 0.78, failing to reach an AUC of 0.8. Using qVA, however, the AUC increases with the number of trials, and an AUC of 0.8 is reached after 5, 11, 19, 34, and 49 trials for changes of 0.1, 0.07, 0.05, 0.03, and 0.02 logMAR, respectively.

4 CONCLUSION

We here introduced qVA, a novel algorithm for visual acuity estimation that uses active learning techniques to quickly adapt to an observer's visual performance, and efficiently only presents informative stimuli to save testing time. A similar strategy has been used before to rapidly estimate the full contrast sensitivity function of an observer by learning the parameters of a parametrized model [14]. Compared to the state of the art, qVA estimates not only the threshold Θ of the psychometric function, but also its range ρ (the inverse of its slope), which has been shown to vary across test populations [4].

Because test-retest repeatability is commonly used in the literature as a proxy for "precision" of a test, we calculated two different repeatability measures. By design, MAP should be less vulnerable to scaling and quantization artefacts when comparing different tests than the standard deviation of test-retest differences. Even though its advantage may not be immediately apparent when looking at the threshold parameter Θ in isolation, Θ and the range ρ had different numerical ranges, and thus yielded different results for MAP and standard deviation. Only with MAP was it possible to evaluate the test-retest repeatability of the twodimensional parameter set of Θ and ρ in combination.

Overall, qVA demonstrated both excellent repeatability and ability to detect change. While the number of trials we simulated here (100 per test run) may be too time-consuming in clinical practice, comparable change detection performance to the current gold standard in clinical trials was achieved after 5–7 rows (15–21 letters vs. 58.9 letters with ETDRS), and performance increased with additional rows. Such improved precision could be used to reduce the number of patients in a clinical trial, which reduces costs and thus in turn would enable to run more trials with e.g. different compounds or different dosing regimes, increasing the chances of achieving clinically meaningful effects. Alternatively, the ability to identify very small differences in intervention effects may enable iterative improvement over current gold standard treatments. For example, the advent of anti-VEGF injections about a decade ago dramatically improved outcomes over then-existing laser treatments for patients with age-related macular degeneration [17], with patients re-gaining more than 0.2 logMAR of acuity instead of a further loss of visual function. However, further improvements so far have proven elusive: Some early studies of new pharmacological compounds, such as the Fovista anti-PGDF therapy, have hinted at additional effect sizes of about 0.02–0.08 logMAR, which are difficult to reliably detect using the ETDRS letter chart. In our simulations, however, 11–49 rows of qVA sufficed to detect such changes (corresponding to a 4.7%–20% change in stimulus size) with an AUC of 0.8. In the future, larger psychophysical studies are needed to empirically validate the potential of the qVA algorithm to rapidly and precisely assess visual function.

Conflict of Interest Declaration

The authors have intellectual property interests in adaptive methods for visual function assessment. LAL also holds employment with Adaptive Sensory Technology, Inc., a company commercializing the techniques presented in this paper.

ACKNOWLEDGMENTS

The authors gratefully acknowledge software implementation support by Simon Thiel and Tiberiu Viuleţ.

REFERENCES

[1] Michael Bach et al. 1996. The Freiburg Visual Acuity Test-automatic measurement of visual acuity. *Optometry and Vision Science* 73, 1 (1996), 49–53.

[2] Michael Biehl. 2017. Biomedical Applications of Prototype Based Classifiers and Relevance Learning. In *International Conference on Algorithms for Computational Biology*. Springer, 3–23.

[3] J. M. Bland and D. G. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1, 8476 (Feb 1986), 307–310.

[4] Andrew Carkeet, Linda Lee, Jennifer R Kerr, and Maile M Keung. 2001. The slope of the psychometric function for Bailey-Lovie letter charts: defocus effects and implications for modeling letter-by-letter scores. *Optometry and Vision Science* 78, 2 (2001), 113–121.

[5] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.

[6] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* 24, 9 (2018), 1342.

[7] Michael Dorr, Tobias Elze, Hui Wang, Zhong-Lin Lu, Peter Bex, and Luis Lesmes. 2018. New precision metrics for contrast sensitivity testing. *IEEE Journal of Biomedical and Health Informatics* 22, 3 (2018), 919–925.

[8] David B Elliott and David Whitaker. 1990. Changes in macular function throughout adulthood. *Documenta Ophthalmologica* 76, 3 (1990), 251–259.

[9] F L Ferris, A Kassoff, G H Bresnick, and I Bailey. 1982. New visual acuity charts for clinical research. *American Journal of Ophthalmology* 94, 1 (1982), 91–96.

[10] Gordon Guyatt, Stephen Walter, and Geoff Norman. 1987. Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases* 40, 2 (1987), 171 – 178. https://doi.org/10.1016/0021-9681(87)90069-5

[11] Karsten U Kortüm, Michael Müller, Christoph Kern, Alexander Babenko, Wolfgang J Mayer, Anselm Kampik, Thomas C Kreutzer, Siegfried Priglinger, and Christoph Hirneiss. 2017. Using Electronic Health Records to Build an Ophthalmologic Data Warehouse and Visualize Patients' Data. *American Journal of Ophthalmology* 178 (2017), 84–93.

[12] Ryan R Kroll, J Gordon Boyd, and David M Maslove. 2016. Accuracy of a wristworn wearable device for monitoring heart rates in hospital inpatients: a prospective observational study. *Journal of Medical Internet Research* 18, 9 (2016).

[13] Janne V Kujala and Tuomas J Lukka. 2006. Bayesian adaptive estimation: The next dimension. *Journal of Mathematical Psychology* 50, 4 (2006), 369–389.

[14] Luis Andres Lesmes, Zhong-Lin Lu, Jongsoo Baek, and Thomas D. Albright. 2010. Bayesian adaptive estimation of the contrast sensitivity function: The quick CSF method. *Journal of Vision* 10, 3 (2010).

[15] Zhong-Lin Lu and Barbara Dosher. 2013. *Visual Psychophysics: From Laboratory to Theory*. MIT Press.

[16] JL Ohlsson and G Villarreal. 2004. Normal visual acuity in 17–18–year olds. *Investigative Ophthalmology & Visual Science* 45, 13 (2004), 4305–4305.

[17] Jan SAG Schouten, Ellen C La Heij, Carroll AB Webers, Igor J Lundqvist, and Fred Hendrikse. 2009. A systematic review on the effect of bevacizumab in exudative age-related macular degeneration. *Graefe's Archive for Clinical and Experimental Ophthalmology* 247, 1 (2009), 1.

[18] R Van Veen, L Talavera Martinez, RV Kogan, SK Meles, D Mudali, JBTM Roerdink, F Massa, M Grazzini, JA Obeso, MC Rodríguez-Oroz, KL Leenders, RJ Renken, JJG De Vries, and M Biehl. 2018. Machine Learning Based Analysis of FDG-PET Image Data for the Diagnosis of Neurodegenerative Diseases. In *Applications of Intelligent Systems 2018*.