New Precision Metrics for Contrast Sensitivity Testing

Michael Dorr, Tobias Elze, Hui Wang, Zhong-Lin Lu, Peter J. Bex, and Luis A. Lesmes

***Abstract*—Visual sensitivity is comprehensively described by the contrast sensitivity function (CSF), but current routine clinical care does not include its assessment because of the time-consuming need to estimate thresholds for a large number of spatial frequencies. The quick CSF method, however, dramatically reduces testing times by using a Bayesian information maximization rule. We evaluate the test–retest variability of a tablet-based quick CSF implementation in a study with 100 subjects who repeatedly assessed their vision with and without optical correction. We first discuss two commonly used measures of repeatability, intraclass correlation and the Bland–Altman Coefficient of Repeatability, and show that they are vulnerable to artifacts. Instead, we propose to formulate precision as an information retrieval task: from all repeat test scores, can we retrieve a certain individual based on their first test score? We then use rank-based analyses such as mean average precision as a better measure to compare different test metrics, and show that the highest test-retest precision is achieved using a summary statistic, the area under the log CSF (AULCSF). This demonstrates the benefit of assessment of the whole CSF compared to sensitivity at individual spatial frequencies only. AULCSF also yields best discrimination performance (99.2%) between measurements that were taken with and without glasses, respectively, even better than CSF Acuity. The tablet-based quick CSF thus enables the rapid and reliable home monitoring of visual function, which has the potential to improve early diagnosis and treatment of ophthalmic pathologies such as diabetic retinopathy or age-related macular degeneration.**

***Index Terms*—Biomarkers, psychometric testing, reproducibility of results.**

Manuscript received November 23, 2016; revised April 18, 2017; accepted May 22, 2017. Date of publication June 25, 2017; date of current version May 3, 2018. This work was supported by the NIH under Grants EY018664, EY021553-01, and EY023902. The work of M. Dorr was supported by the Bavarian State Ministry for Research and Education's Elite Network Bavaria. The work of H. Wang was supported by a China Scholarship Council Grant. *(Corresponding author: Michael Dorr.)*

M. Dorr is with the Institute for Human-Machine Communication, Technische Universität München, Munich 80333, Germany (e-mail: michael.dorr@tum.de).

T. Elze is with the Schepens Eye Research Institute, Massachusetts Eye and Ear Infirmary, Boston, MA 02114 USA (e-mail: tobias-elze@tobias-elze.de).

H. Wang is with the Jilin University of Finance and Economics, Changchun 3699, China (e-mail: huiwangedu@163.com).

Z.-L. Lu is with the Center for Cognitive and Behavioral Brain Imaging, Arts and Sciences, Ohio State University, Columbus, OH 43210 USA (e-mail: lu.535@osu.edu).

P. J. Bex is with the College of Science, Northeastern University, Boston, MA 02115 USA (e-mail: p.bex@northeastern.edu).

L. A. Lesmes is with Adaptive Sensory Technology, San Diego, CA 92121, USA (e-mail: luis.lesmes@adaptivesensorytech.com).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2017.2708745

I. Introduction

VISION is the most important sense for many everyday tasks, and blindness ranks highly among most-feared ailments [1]. Because of the global demographic trends of aging and obesity, however, the prevalence of vision loss is predicted to rise dramatically over the next few decades (e.g., 135% by 2050 according to [2]). The most common causes for vision loss are neuro-degenerative eye diseases such as age-related macular degeneration [3], diabetic retinopathy [4], or glaucoma [5], for which early diagnosis and treatment are critical. While moderate levels of myopia, or nearsightedness, can typically be corrected by optical devices or surgery, high myopia is also associated with severe eye conditions such as glaucoma and retinal detachment [6]; prevalence of myopia is estimated between 15 and almost 50% in the global adult population [7]. Therefore, regular monitoring of visual function at least in at-risk populations would be desirable; however, the economic costs and practical burdens of doctor's office visits lead to poor compliance. For example, less than one third of patients returned for a diabetic retinopathy screening at least every 15 months in a recent study, with geographic access as a significant predictor of compliance [8]. As a consequence, a number of vision tests have been developed that can be run at home without supervision. In this paper, we are introducing new metrics to evaluate precision and repeatability of clinical measurements, and thus their suitability for disease monitoring.

For example, several apps have been developed that utilize high-resolution screens of smartphones and tablets to assess visual function [9]–[12], but they are typically restricted to acuity-like measures, i.e., the highest resolving power of the optical system to discern black-on-white optotypes; the Amsler grid [13] has also been implemented as an app [14], and hardware add-ons are available to determine the eye's refractive error on a smartphone [15].

However, visual function depends not only on optical, but also neural factors. The contrast sensitivity function (CSF) thus provides a more fundamental and comprehensive assessment of visual performance than acuity by relating spatial frequency, or size, to the minimum contrast required to discern patterns of that size. Contrast sensitivity is better correlated with visual quality of life [16], [17] and notably may be impaired in neurodegenerative ocular pathologies even when acuity is unaffected [18], [19]. Despite its clinical value, contrast sensitivity testing has not found its way into routine clinical care because of practical constraints. The straightforward approach of testing all possible combinations of spatial frequency and contrast is

too time-consuming, and coarse, heuristic sampling of this two-dimensional parameter space [20] limits a test's sensitivity. A further constraint is imposed by the prevalent use of paper charts that may be memorized, and only recently have contrast sensitivity tests begun to utilize high-fidelity computer and tablet displays [21]–[25]. A more principled approach to efficiently assess the whole CSF has been proposed in [26], based on the insight that the human CSF can be described by only four parameters [27]. By using a Bayesian approach that maximizes information gain over a very large set of possible stimuli, the number of trials to reliably estimate the CSF was reduced from several hundreds using traditional methods to several dozens using the quick CSF method.

Because of the additional information provided by the CSF compared to scalar measures such as acuity, a variety of features can be derived from a single test, and the output of other tests as well might differ in sensitivity, quantization, and dynamic range. Thus, the comparative evaluation of their precision and repeatability is challenging. In the following, we will analyze established methods of the assessment of repeatability and introduce rank-based analysis as a better measure of clinical test-retest variability. For an empirical evaluation, we use a tablet-based implementation of the quick CSF [22] to collect a large data set of vision in myopic (near-sighted) subjects and controls.

II. Precision and Repeatability of Measurements

In order to be clinically useful, measurements need to be precise and sensitive to changes in the ground truth due to disease progression or remediation. Two major sources of imprecision typically are noise in the measurement device and moment-to-moment variability of the phenomenon under observation; in the case of the contrast sensitivity function, such variability may occur because of neural noise, lack of attention, finger lapses, and small diurnal changes in visual sensitivity. According to ISO norm 5725-1, precision is the general term for similarity between repeated measurements [28]. For clinical testing, the standard tools to assess this similarity, or test-retest variability, are the intra-class correlation coefficient (ICC) and the Bland-Altman coefficient of repeatability (CoR) [29], which in line with the ISO definition is based on the standard deviation of test-retest differences.

However, precision in this sense may be achieved almost trivially by making a clinical test insensitive to change, for example by strong quantization of the test scores. In the extreme case, a test with a binary outcome, such as light perception, might have almost perfect repeatability but little discriminatory power for changes in visual function, or between different members of the population. In the following, we will therefore use the term rank precision to denote the desired quality of a measurement system that i) returns (almost) the same output for repeated measurements of the same ground truth; and ii) returns (with a high probability) different outputs for different ground truths.

The common repeatability metrics have further drawbacks. The ICC is dominated by the values at either end of the test range and is therefore sensitive to outliers; the CoR does not suffer from this problem, and also may provide an intuitive threshold for how much change between two tests should be considered statistically significant. However, this threshold rests on the assumption that tests are homoscedastic, i.e., that measurement error is independent of the magnitude of the ground truth (e.g., patients with poor vision perform tests as reliably as normal-sighted controls); its usefulness is also limited by the quantization of many tests. For example, the popular ETDRS vision chart has a coarse resolution of 0.1 log10 units in optotype sizes; while the 95% Coefficient of Repeatability may be 0.14 log10 units of test-retest difference [30], an individual patient's visual acuity needs to change at least 0.2 log10 units for reliable change detection.

Finally, as noted by Bland and Altman already, absolute CoR values do not directly relate to clinical meaningfulness. In the following, we will see that measures with higher ('worse') coefficients of repeatability may yet be more precise than measures with a lower CoR. *Responsiveness*, which is the ratio of the clinically relevant difference to variability in stable subjects, has been proposed to address this issue [31]; however, responsiveness assumes normally distributed data and, like CoR, a measurement error that is independent of the magnitude of the ground truth.

For these reasons, we choose to evaluate the quick CSF method using techniques from information retrieval. In information retrieval, the goal is to retrieve, from a large set of documents, only relevant items following a query. In our case, the query is the first measurement $M_1^{s_i}$ obtained from a specific subject s_i, and the one relevant item that we want to retrieve from the set of all second measurements M_2^S across all subjects is the second measurement $M_2^{s_i}$. The M_2^S are sorted by their similarity to $M_1^{s_i}$, and Mean Average Precision (MAP) as a standard tool to assess ranked retrieval results can be used to evaluate test-retest rank precision. Because MAP scores may not be immediately intuitive, we also introduce a new rank-based measure, Fractional Rank Precision (FRP), that has an intuitive dynamic range from 0.5 (chance) to 1.0 (perfect test-retest identification).

III. Methods

A. iPad-Based Quick CSF Implementation

We used an iOS-based implementation of the quick CSF method [22], [26] with several changes to experimental parameters. Notably, while previous implementations had used sine-wave gratings as stimuli, we here used a set of 10 Sloan letters [32] that were bandpass-filtered with a raised cosine window with peak frequency 4.5 cycles per letter to increase spatial frequency specificity [33]; the reduced guessing rate when going from two to 10 alternatives has been shown to result in greater efficiency [34].

Comfortably seated, subjects held an iPad 4 that was set to a mean screen luminance of 185 cd/m^2 at a viewing distance of 60 cm. In each of 50 trials, a bandpass-filtered Sloan letter was shown for 500 ms. Its identity was chosen at random, and its spatial frequency and contrast were chosen as below. After stimulus presentation, the subject registered their response by touch-selecting a letter from an array on the screen (10-Alternatives

Forced Choice). The stimulus space was log-spaced and comprised 24 spatial scales (0.64 to 41 cycles per degree (cpd)) and 48 contrast levels (0.2 to 100%); for this set of 1152 stimulus parameter options, the quick CSF method determined the expected information gain based on the history of trials and a sample of possible CSFs in a four-dimensional search space (describing peak spatial frequency $f_{\max}$ and peak sensitivity $\gamma_{\max}$, bandwidth β, and a low-frequency truncation parameter δ). In order to avoid uninformative regions of the stimulus space, the next stimulus was then chosen to maximize information gain.

B. Experimental Data Collection

Participants were recruited among students and faculty of Jilin University of Finance and Economics; the experimental design followed the principles of the Declaration of Helsinki and was approved by the university's ethics committee under protocol no. IPBWH1101.

Complete data sets as described below were collected from 100 subjects that were classified as either *myopes* (self-reporting as wearing optical correction; 62 participants aged 18–75 years; mean = 22, s.d. = 5.7 years; self-reported refraction mean = −3.9D, s.d. = 1.6) or *controls* (no optical correction; 38 participants, 17–45 years; mean = 22, s.d. = 7.8 years). Six additional subjects participated in the study, but were later excluded because of missing data (the first 5 subjects were tested only monocularly; one subject did not complete 50 trials per test).

All subjects ran the test multiple times. Controls were tested monocularly in both eyes and binocularly; to estimate the test-retest variability of the procedure, one of these three conditions (randomly chosen) was repeated. Myopes were tested both with and without their optical correction, with one repeat, for a total of 7 conditions.

C. Statistical Analysis

For improved analysis accuracy, data were re-scored on a regular workstation PC to overcome limitations of computational power and memory on the iPad. Whereas the iPad version used a CSF search space with about 300000 nodes, re-scoring utilized ca. 2 million grid nodes. Furthermore, the quick CSF assumed a uniform prior during the original experiment. During off-line analysis, the posterior distribution after 50 trials for all of the 586 data sets available was used as the prior for an additional re-scoring step.

In principle, the test outcome is completely described by a joint probability distribution of the four CSF parameters $f_{\max}$, $\gamma_{\max}$, β, δ. For better intuition, however, we sampled 2500 CSF parameter sets from the posterior distribution and computed the median contrast sensitivity for 2000 spatial frequencies in the range from 0.36 to 73 cpd (extending the range of presentable stimuli by 0.25 log10 units on either side, step size 0.0011 log10 units). From this, we picked sensitivities for six individual spatial frequencies (1, 1.5, 3, 6, 12, and 18 cpd) and also computed CSF Acuity, i.e., the intersection of the CSF with the x axis (where contrast threshold is 100%). Furthermore, we calculated the AULCSF [35], i.e., the area under the log CSF. Instead of computing the area for the entire range of spatial frequencies that we tested (0.64 to 41 cpd), which would make it impossible to compare this number across studies that differ in, e.g., screen size or viewing distance, we integrated over the range from 1.5 to 18 cpd. This range is in line with guidelines by the Food and Drug Administration (FDA) and, e.g., ANSI standards for evaluation of multifocal intraocular lenses [36], which recommend testing contrast vision at 1.5, 3, 6, and 12 cpd for mesopic and at 3, 6, 12, and 18 cpd under photopic conditions. Specifically, we used Riemannian integration with a step size of 0.0011 log10 units (see above) on the x axis of the CSF plotted as logarithmic sensitivity (inverse of threshold contrast) against logarithmic spatial frequency. Because the range of 1.5 to 18 cpd spans close to one log10 unit (1.08), our AULCSF approximately corresponds to the mean sensitivity over this range.

For each of the features above, we performed rank-precision analysis as follows: Let M_2^S be the set of all N re-test measurements. We sort this set by distance to the first measurement $M_1^{s_i}$ of one specific subject s_i in ascending order, and use $r(i)$ to denote the rank of $M_2^{s_i}$ in the sorted sequence. Because we have exactly one relevant item (namely $M_2^{s_i}$) per query, average precision can be simply computed by AveP $= 1/r(i)$, and we obtain mean AveP (MAP) [37] by repeating this process for each subject S_k, $k = 1, \ldots, N$; for additional robustness, we computed MAP both for identifying tests from retests and for identifying retests from tests.

In principle, the dynamic range of MAP extends from near-zero $(2/N)$ for very poor rank precision to 1.0 for perfect rank precision and thus provides a very intuitive description of performance. In practice, however, greater emphasis is placed on very good query matches (for example, a pair of queries with returned ranks of 1 and 3 yields a higher MAP than returned ranks of 2 and 2), and the meaningfulness of a MAP difference may be less clear. Therefore, we propose a new measure, also based on ranking retests by their similarity to the tests, that we call Fractional Rank Precision

$$\text{FRP} = \frac{1}{N} \sum_i \frac{N - (r(i) - 1)}{N}.$$

This measure ranges from 0.5 (chance) to 1.0 (perfect test-retest identification) and has an intuitive interpretation because FRP describes test-retest variability in terms of population variability: if we were to sort our subjects by visual function, on average the percentile for a specific subject in repeated assessments would differ by (1-FRP)/2.

Repeatability measures are affected by quantization of test results, and the size of this effect depends on how close the observed values lie to the quantization thresholds. For example, because of measurement noise, a hypothetical test might yield test-retest measures of 1.06 and 0.99; a test with 0.1 units step size might then report 1.1 and 1.0, respectively. For measures of 1.03 and 0.96, however, the same actual test-retest difference would yield a quantized output of 1.0 for both test and retest, and thus seemingly show perfect repeatability. In order to assess this effect for our analysis on quantization (Section IV-B), we resampled our data set 1000 times, adding Gaussian noise to each test with a standard deviation that mimicked the distribution of observed test-retest differences.

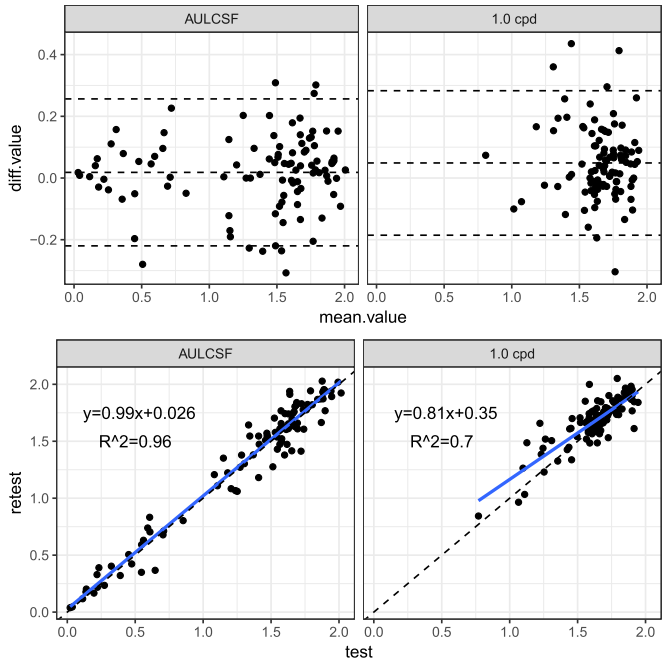

Fig. 1. Top row, Bland-Altman plots for AULCSF and sensitivity at 1.0 cpd. Dashed lines indicate bias and 95% Limits of Agreement for test-retest differences. Bottom row, correlation plots with linear regression lines.

D. Prediction of Optical Correction Status

To further assess the usefulness of different features in describing functional vision, we performed a simple classification task. For each subject with optical correction, we had four pairs of measurements that differed in their correction status (left eye, right eye, both eyes, and one repeat), i.e., 248 corrected-uncorrected pairs overall. Under the assumption that optical correction improves functional vision, we calculated the number of times a feature was higher (better) in the measurement with correction. In order to assess the effect size of optical correction for a feature, relative to its test-retest variability, we also looked at the 124 pair-wise comparisons based on triplets of test-retest pairs and their one correction-complement measurement (e.g., for a repeated binocular measurement with glasses, the correction-complement was the binocular measurement without glasses): how often was the difference between corrected and uncorrected measurement greater than the difference between the repeated measurements (in our data, 30 uncorrected test-retest pairs and 32 corrected)?

IV. Results

A. Test-Retest Variability and Rank Precision

Bland-Altman CoR and inter-class correlation for two selected features are graphically shown in Fig. 1 and full results are tabularized in Table I. As can be seen in Fig. 1, test results for sensitivity at 1.0 cpd range from roughly 0.8 to 2.0, whereas AULCSF scores range down to almost zero. In both top panels, the plots do not show obvious dependencies of test-retest differences on the magnitude of the mean test score; linear regression of absolute differences against the mean confirms no statistically significant relationship (for AULCSF, $p = 0.43$, $R^2 < 0.01$; for sensitivity at 1 cpd, $p = 0.14$, $R^2 = 0.02$). The standard deviation of differences for 1.0 cpd is slightly smaller than for AULCSF. Scatterplots in the bottom panels also show the strong linear relationship between test and retest scores. For the AULCSF feature, the test-retest agreement is excellent, with an explained variance R^2 between test and retest of 0.96; the slope of the regression line is 0.994 ($p << 0.001$) with a non-significant intercept term ($p = 0.4$).

TABLE I
Numerical Results for Different Test Features and Test-Retest Variability Measures

Feature	ICC	CoR	MAP	FRP
AULCSF	**0.977**	0.238	**0.231**	**0.865**
CSF Acuity	0.966	**0.197**	0.188	0.839
1 cpd	0.838	0.234	0.132	0.748
1.5 cpd	0.928	0.225	0.161	0.797
3 cpd	0.965	0.283	0.203	0.829
6 cpd	0.973	0.301	0.173	0.842
12 cpd	0.957	0.329	0.117	0.837
18 cpd	0.912	0.396	0.122	0.816

Best scores are highlighted in bold.

MAP scores are listed in the fourth column of Table I. The AULCSF yields highest MAP; i.e., the AULCSF is the best feature to identify a pair of test-retest measurements, even though it summarizes over a broad range of spatial frequencies. Sensitivities for the lower spatial frequencies 1.0 and 1.5 cpd vary less across subjects in our study and are therefore less discriminative. Despite their worse rank precision, however, these sensitivities show smaller test-retest CoR than the AULCSF. The same pattern can be seen for fractional rank precision in the fifth column of Table I; again, AULCSF is the feature with the highest test-retest precision. Because of the more linear nature of FRP compared to MAP, however, the worst feature according to FRP is sensitivity at 1 cpd, which has relatively small dynamic range; for MAP, the relatively large number of ties at higher spatial frequencies (many subjects scoring 0) makes sensitivities at 12 and 18 cpd the worst features instead.

B. Quantization Artifacts

Fig. 2 shows how the repeatability measures for the feature AULCSF change as a function of test score quantization. For extreme quantization levels (all test-retest scores are the same), CoR even indicates perfect repeatability; ICC is ill-defined (division by zero), and only MAP and FRP give scores in line with intuition ($2/N$ and 0.5, respectively, i.e., chance). Intuitively, a measure of precision should also monotonically decrease with increasing quantization; for CoR and ICC, this is only the case for the average of 1000 resamplings of our data set shown in Fig. 2. The variability across resamplings of our data set with 100 pairs of repeated measurements is shown in Fig. 3: because ICC and CoR are sensitive to small perturbations of the data, monotonicity often does not hold.

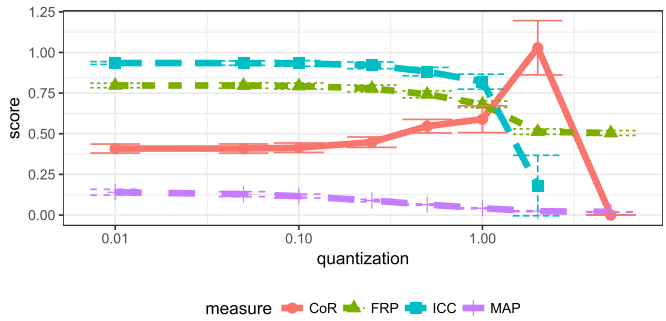

Fig. 2. Effect of quantization of AULCSF scores on the repeatability measures Coefficient of Repeatability (COR), intra-class correlation (ICC), Mean Average Precision (MAP), and Fractional Rank Precision (FRP). For extreme quantization, the CoR goes to 0 (perfect repeatability) and ICC becomes ill-defined. Error bars indicate +/− s.d. of 1000 resamplings of our empirical data, where noise was added to our measurements with the same characteristics as the observed test-retest difference distribution.

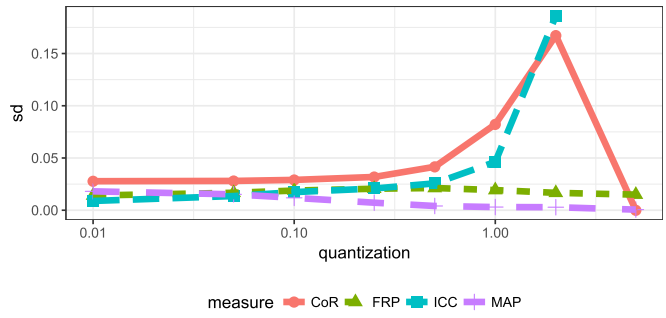

Fig. 3. Effect of quantization of AULCSF scores on the variability of repeatability measures; plotted values correspond to the magnitude of the error bars in Fig. 2. At higher quantization levels, CoR and ICC are more sensitive to random perturbations of the data than FRP and MAP.

TABLE II
HOW OFTEN CORRECTED VISION SCORES ARE HIGHER THAN UNCORRECTED VISION (MIDDLE COLUMN), AND HOW OFTEN THIS DIFFERENCE IS GREATER THAN THE TEST-RETEST DIFFERENCE (RIGHT)

Feature	Optical status prediction	Correction effect > test-retest diff
AULCSF	**0.992**	**0.984**
CSF Acuity	0.976	0.935
1 cpd	0.847	0.564
1.5 cpd	0.887	0.806
3 cpd	0.968	0.887
6 cpd	**0.992**	**0.984**
12 cpd	**0.992**	**0.984**
18 cpd	0.952	0.855

C. Prediction of Optical Correction Status

Results for the optical correction status are shown in the middle column of Table II. Even though the milder levels of myopia in our cohort should not impair vision substantially at the short viewing distance of 60 cm, all features correlate well with functional vision, but to a varying degree. For all but two measurement pairs, the AULCSF was higher in measurements with glasses than without glasses, yielding an accuracy of 99.2%. CSF Acuity performed only slightly worse with an accuracy of 97.6%; sensitivity to very low spatial frequencies (1.0 cpd) changed less with optical correction and thus yielded 84.7%. The rightmost column of Table II shows how often the test difference due to optical correction was greater than the difference between the corresponding test-retest pair. As with optical status prediction, AULCSF and sensitivities at 6 and 12 cpd perform best, but sensitivity at 1 cpd is particularly bad with a small effect size relative to test-retest variability.

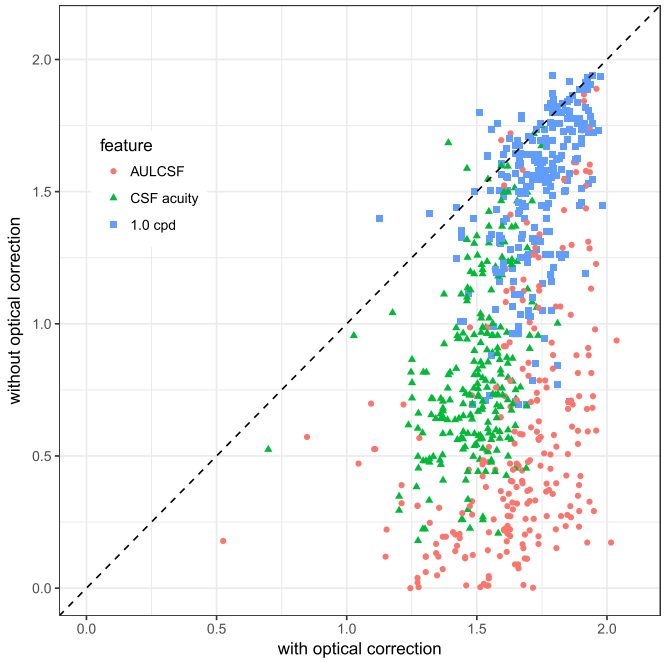

Fig. 4. Visualization of the effect of optical correction on AULCSF, CSF Acuity, and sensitivity at 1 cpd. The AULCSF shows best separation between the two conditions (cf. Table II).

This effect can also be seen in Fig. 4: AULCSF, CSF Acuity, and sensitivity at 1 cpd all are clearly biased towards better scores with correction, but the 1 cpd score is clustered much more tightly around the diagonal (i.e., less discriminative).

V. DISCUSSION

Novel vision tests rest on computationally expensive algorithms to provide a more comprehensive description of visual function. Thus, they have the potential to improve clinical care and research by more precise measurement of the effects of disease progression or ophthalmic interventions. A common proxy to study precision of a test is to assess its test-retest variability; however, standard methods to assess clinical precision and test-retest variability may be inadequate for the more complex, higher-dimensional test outputs.

By using a Bayesian information maximization algorithm, the quick CSF [26] efficiently estimates the whole contrast sensitivity function within 2–5 minutes. Here, we extended a tablet-based implementation of the quick CSF method [22] to use a greater number of optotypes and thus increase efficiency even further. Various other apps exist that enable vision testing on smartphones and tablets [38]; however, these typically only transfer existing, hand-scored eye chart technology onto

electronic devices, without necessarily utilizing the greater computational power these devices offer.

In the present study, the quick CSF method was always initialized with a flat prior during data collection. In principle and in the spirit of personalized medicine, even higher efficiency may be obtained by using the posterior of previous tests by the same subject as the prior, which would reduce the number of initial trials where the algorithm homes in on the general threshold area. At an intermediate level of personalization, we here used the posterior of all subjects as a prior during rescoring. Given a sufficiently large data set, it may also be possible to use a prior that is representative of the entire population.

Ultimately, the purpose of a vision test is to reliably detect even small changes in visual function. High repeatability is only a necessary but not sufficient condition for this goal; for example, test score quantization such as in common paper charts may improve repeatability, but impairs the ability to track change. Because commonly used measures of test-retest variability suffer from such artifacts and are vulnerable to outliers, we here proposed to use rank-based measures that express test-retest variability in terms of population variability. Using mean average precision and fractional rank precision, we compared different features of the quick CSF method and found that the summary statistic AULCSF was more rank-precise in identifying pairs of test-retest measurements than, e.g., sensitivity at low spatial frequencies, despite a higher Bland-Altman Coefficient of Repeatability. Rank-based measures can also be used to compare different vision tests, although it should be noted that scores depend on the specific subject cohort and thus can be meaningfully compared only within studies. A recent study found that the AULCSF of the quick CSF had a similar CoR as the well-established Pelli-Robson contrast sensitivity chart and the ETDRS visual acuity chart in a normative cohort, but substantially better FRP [39]. The FRP of the AULCSF was reduced to that of Pelli-Robson and ETDRS only when the AULCSF scores (that ranged from about 0.5 to 2 log10 units) were strongly quantized to a step size of 0.25 log10 units.

We also showed that the quick CSF method was sensitive to the change in visual function due to optical correction; in only two out of 248 measurement pairs was the AULCSF for a test with glasses lower than the corresponding test without. In contrast, the Pelli-Robson chart reliably detects defocus in myopes only at -3D or more refractive power [40] (in our data set, 27 out of 124 study eyes required less than 3 diopters of correction). However, one limitation of our study is that refractive error was only self-reported.

Our tablet-based quick CSF implementation has already been used for clinical and neuroscientific research both inside and outside of the laboratory, e.g., to screen amblyopia in children [41] or to track the development of vision in congenitally blind children and teenagers after cataract removal [42]. Yet, in these studies, the test was still supervised by a researcher. The greatest benefit of a precise vision test on a portable device might be realized in regular home monitoring; this is particularly true for neurodegenerative diseases such as age-related macular degeneration and diabetic retinopathy, where early detection of subtle changes in visual function might enable interventions to stop or even revert vision loss. However, completely unsupervised settings such as the home pose further challenges to the robustness of a test; for example, different light conditions and glare might affect test results. Changes in viewing distance also change the retinal size of presented stimuli, but we note that the front-facing camera of the iPad may be used to ensure viewing distance compliance.

In summary, we here made theoretical and practical contributions to the comparative analysis of clinical outcome measures, and demonstrated the sensitivity and precision of a vision test on a portable device.

Acknowledgement

The authors would like to thank J. Mulligan and an anonymous reviewer for their constructive feedback on a first version of the manuscript.

M. Dorr, Z.-L. Lu, P. J. Bex, and L. A. Lesmes declare an intellectual property and financial interest in Adaptive Sensory Technology (AST), a company currently commercializing technology related to the research presented here. H. Wang declares having received travel support from AST to present these results at a conference.

References

[1] J. f. Jorkasky, "Attitudinal survey of minority populations on eye and vision health and research: Research america national public opinion poll," Aug. 2014. [Online]. Available: http://www.researchamerica.org/sites/default/files/uploads/AEVRRApoll.pdf

[2] J. S. Wittenborn and D. B. Rein, "The future of vision: Forecasting the prevalence and cost of vision problems," NORC at the University of Chicago. Prepared for Prevent Blindness, Chicago, IL, USA, Tech. Rep., 2014. [Online]. Available: http://forecasting.preventblindness.org

[3] A. Chopdar, U. Chakravarthy, and D. Verma, "Age related macular degeneration," *Brit. Med. J.*, vol. 326, no. 7387, pp. 485–488, 2003.

[4] P. S. Silva, J. D. Cavallerano, L. M. Aiello, and L. P. Aiello, "Telemedicine and diabetic retinopathy: Moving beyond retinal screening," *Arch. Ophthalmol.*, vol. 129, no. 2, pp. 236–242, Feb. 2011.

[5] H. A. Quigley and A. T. Broman, "The number of people with glaucoma worldwide in 2010 and 2020," *Brit. J. Ophthalmol.*, vol. 90, no. 3, pp. 262–267, 2006.

[6] P. J. Foster and Y. Jiang, "Epidemiology of myopia," *Eye*, vol. 28, no. 2, pp. 202–208, 2014.

[7] C. W. Pan, D. Ramamurthy, and S. M. Saw, "Worldwide prevalence and risk factors for myopia," *Ophthalmic Physiol. Opt.*, vol. 32, no. 1, pp. 3–16, 2012.

[8] D. J. Lee *et al.*, "Dilated eye examination screening guideline compliance among patients with diabetes without a diabetic retinopathy diagnosis: The role of geographic access," *BMJ Open Diabetes Res. Care*, vol. 2, no. 1, 2014, Art. no. e000031.

[9] J. M. Black *et al.*, "An assessment of the iPad as a testing platform for distance visual acuity in adults," *BMJ Open*, vol. 3, no. 6, 2013, Art. no. e002730.

[10] P. K. Kaiser, Y.-Z. Wang, Y.-G. He, A. Weisberger, S. Wolf, and C. H. Smith, "Feasibility of a novel remote daily monitoring system for age-related macular degeneration using mobile handheld devices: Results of a pilot study," *Retina*, vol. 33, no. 9, pp. 1863–1870, 2013.

[11] Z.-T. Zhang, S.-C. Zhang, X.-G. Huang, and L.-Y. Liang, "A pilot trial of the iPad tablet computer as a portable device for visual acuity testing," *J. Telemed. Telecare*, vol. 19, no. 1, pp. 55–59, 2013.

[12] P. A. Gounder, E. Cole, S. Colley, and D. M. Hille, "Validation of a portable electronic visual acuity system," *J. Mobile Technol. Med.*, vol. 3, no. 2, pp. 35–39, 2014.

[13] M. F. Marmor, "A brief history of macular grids: From Thomas Reid to Edvard Munch and Marc Amsler," *Survey Ophthalmol.*, vol. 44, no. 4, pp. 343–353, 2000.

[14] S. Luk, K. Chen, and N. Davies, "Variation of online Amsler grid from mobile apps, YouTube and Google," *Acta Ophthalmologica*, vol. 91, 2013.
[15] V. F. Pamplona, A. Mohan, M. M. Oliveira, and R. Raskar, "NETRA: Interactive display for estimating refractive errors and focal range," *ACM Trans. Graph.*, vol. 29, no. 4, 2010, Art no. 77.
[16] C. Owsley, "Contrast sensitivity," *Ophthalmol. Clin. North Amer.*, vol. 16, no. 2, pp. 171–177, Jun. 2003.
[17] J. P. Stellmann, K. L. Young, J. Pöttgen, M. Dorr, and C. Heesen, "Introducing a new method to assess vision: Computer-adaptive contrast-sensitivity testing predicts visual functioning better than charts in multiple sclerosis patients," *Multiple Sclerosis J., Exp., Transl. Clin.*, vol. 1, pp. 1–8, 2015.
[18] L. F. Jindra and V. Zemon, "Contrast sensitivity testing: A more complete assessment of vision," *J. Cataract Refractive Surg.*, vol. 15, no. 2, pp. 141–148, Mar. 1989.
[19] R. L. Woods and J. M. Wood, "The role of contrast sensitivity charts and contrast letter charts in clinical practice," *Clin. Exp. Optometry*, vol. 78, no. 2, pp. 43–57, 1995.
[20] K. Thayaparan, M. D. Crossland, and G. S. Rubin, "Clinical assessment of two new contrast sensitivity charts," *Brit. J. Ophthalmol.*, vol. 91, no. 6, pp. 749–752, 2007.
[21] J. B. Mulligan, "A method for rapid measurement of contrast sensitivity on mobile touch-screens," in *Proc. Human Vis. Electron. Imag.* (ser. Proc. SPIE), 2016, pp. HVEI–104.1–HVEI–104.6.
[22] M. Dorr, L. Lesmes, Z.-L. Lu, and P. Bex, "Rapid and reliable assessment of the contrast sensitivity function on an iPad," *Investigative Ophthalmol. Vis. Sci.*, vol. 54, pp. 7266–7273, 2013.
[23] T. M. Aslam, I. J. Murray, M. Y. T. Lai, E. Linton, H. J. Tahir, and N. R. A. Parry, "An assessment of a modern touch-screen tablet computer with reference to core physical characteristics necessary for clinical vision testing," *J. Roy. Soc. Interface*, vol. 10, no. 84, 2013, Art. no. 20130239.
[24] A. Turpin, D. J. Lawson, and A. M. McKendrick, "PsyPad: A platform for visual psychophysics on the iPad," *J. Vis.*, vol. 14, no. 3, 2014, Art. no. 16.
[25] M. Rodríguez-Vallejo, L. Remón, J. A. Monsoriu, and W. D. Furlan, "Designing a new test for contrast sensitivity measurement with iPad," *J. Optometry*, vol. 8, no. 2, pp. 101–108, 2015.
[26] L. A. Lesmes, Z.-L. Lu, J. Baek, and T. D. Albright, "Bayesian adaptive estimation of the contrast sensitivity function: The quick CSF method," *J. Vis.*, vol. 10, no. 3, pp. 1–21, 2010.
[27] A. M. Rohaly and C. Owsley, "Modeling the contrast-sensitivity functions of older adults," *J. Opt. Soc. Amer. A*, vol. 10, no. 7, pp. 1591–1599, Jul. 1993.
[28] *Accuracy (Trueness and Precision) of Measurement Methods and Results—Part 1: General Principles and Definitions* , ISO 5725-1:1994, 1994.
[29] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 1, no. 8476, pp. 307–310, Feb. 1986.
[30] R. Beck *et al.*, "A computerized method of visual acuity testing: Adaptation of the early treatment of diabetic retinopathy study testing protocol," *Amer. J. Ophthalmol.*, vol. 135, no. 2, pp. 194–205, 2003.
[31] G. Guyatt, S. Walter, and G. Norman, "Measuring change over time: Assessing the usefulness of evaluative instruments," *J. Chronic Diseases*, vol. 40, no. 2, pp. 171–178, 1987.
[32] L. L. Sloan, "New test chart for the measurement of visual acuity at far and near distance," *Amer. J. Ophthalmol.*, vol. 48, pp. 807–813, 1959.
[33] J. J. McAnany and K. R. Alexander, "Contrast sensitivity for letter optotypes vs. gratings under conditions biased toward parvocellular and magnocellular pathways," *Vis. Res.*, vol. 46, no. 10, pp. 1574–1584, 2006.
[34] F. Hou, L. Lesmes, P. Bex, M. Dorr, and Z.-L. Lu, "Using 10AFC to further improve the efficiency of the quick CSF method," *J. Vis.*, vol. 15, no. 2, pp. 1–18, 2015.
[35] R. Applegate, G. Hilmantel, and H. Howland, "Area under the log contrast sensitivity function: A concise method of following changes in visual performance," *OSA Tech. Dig. Series*, vol. 1, pp. 98–101, 1997.
[36] A. N. S. I. C. Z80, *American National Standard for Ophthalmics: Multifocal Intraocular Lenses*. Fairfax, VA, USA: Opt. Lab. Assoc., 2007.
[37] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[38] A. S. M. Mosa, I. Yoo, and L. Sheets, "A systematic review of healthcare applications for smartphones," *BMC Med. Informat. Decision Making*, vol. 12, no. 1, Jul. 2012, Art. no. 67.
[39] L. Lesmes, A. Bittner, Z.-L. Lu, P. Bex, and M. Dorr, "Distinguishing the contribution of precision and repeatability to vision testing," in *Proc. Assoc. Res. Vis. Ophthalmol. Abstracts*, 2017, paper 2204 - A0342.
[40] A. Bradley, J. Hook, and J. Haeseker, "A comparison of clinical acuity and contrast sensitivity charts: Effect of uncorrected myopia," *Ophthalmic Physiol. Opt.*, vol. 11, no. 3, pp. 218–226, 1991.
[41] M. Kwon *et al.*, "Rapid assessment of core visual deficits in amblyopia," *Investigative Ophthalmol. Vis. Sci. (Association for Research in Vision and Ophthalmology Annu. Meet. Abstracts)*, vol. 54, Jun. 2013, Art no. 2663.
[42] A. Kalia *et al.*, "Development of pattern vision following early and extended blindness," *Proc. Nat. Acad. Sci.*, vol. 111, no. 5, pp. 2035–2039, 2014.