Reliability of the compensation comparison stray-light measurement method

Joris E. Coppens
Luuk Franssen
Netherlands Ophthalmic Research Institute
Amsterdam, Netherlands
E-mail: j.coppens@ioi.knaw.nl

L. J. van Rijn
Vrije Universiteit Medical Centre
Amsterdam, Netherlands

Thomas J. T. P. van den Berg
Netherlands Ophthalmic Research Institute
Amsterdam, Netherlands

Abstract. The compensation comparison (CC) method is a psychophysical technique to measure retinal stray light. It uses a two alternative forced choice (2AFC) measurement paradigm. The 25 binary (0 and 1) responses resulting from the 2AFC test are analyzed using maximum likelihood estimates. The likelihood function is used to give two quantities: the most likely stray-light level of the eye under investigation, and the accuracy of this estimate [called expected standard deviation (ESD)]. The CC method is used in 2422 subjects of the GLARE study. Each eye is tested twice to allow analysis of measurement repeatability. Furthermore, the large amount of responses is used to evaluate the shape of the psychometric function, for which a mathematical model is used. The shape of the psychometric function found by averaging the 0 and 1 responses fit well to the model function. Data sorted according to ESD show differences in the shape of the psychometric function between good and bad observers. These different shapes for the psychometric function are used to reanalyze the data, but the stray-light results remain virtually identical. ESD proves to be an efficient tool to detect unreliable measurements. In clinical practice, ESD may be used to decide whether to repeat a measurement. [DOI: 10.1117/1.2209555]

Keywords: glare; image quality; opthalmology; scattering; stray light; vision.

Paper 05273R received Sep. 23, 2005; revised manuscript received Jan. 3, 2006; accepted for publication Jan. 4, 2006; published online Jun. 7, 2006.

1 Introduction

Intraocular light scatter is the phenomenon where part of the light reaching the retina does not partake in normal image formation.[1] Most rays originating from a certain point in space are converged by the refracting elements of the eye to the focal spot on the retina. Some of the rays, however, are dispersed to other areas by optical imperfections of the eye. This already occurs in the healthy eye,[2] but to a much larger extent in pathological states, such as cataract and corneal dystrophy.[3] These dispersed rays are distributed all over the retina, but with decreasing densities at distances farther away from the original focal spot. The luminance distribution on the retina of an eye looking at a point source is called the point spread function (PSF). The large angle part of this PSF (angles from 1 to 90 deg) is called retinal stray light. Due to stray light, the retinal light distribution in any visual environment is composed of two parts: the image of the external world based on the more or less properly focused rays, superimposed on a background caused by the dispersed rays. As a result, contrast is reduced in the retinal image. The severity of this contrast reduction depends on the luminance ratio between background and image. This ratio is a function of the optical clarity of the eye, and can be quantified and expressed in the physically well-defined retinal stray-light parameter s.[1,4]

The extreme situation of contrast loss due to intraocular light scatter is represented by the classical glare condition:[2] strong light somewhere in the visual field when a weakly lit object has to be observed. In such a situation, the contrast of the retinal image may drop below the contrast threshold, and can lead to complete blinding. A typical situation is blinding by oncoming traffic at night.

Recently, a new test for measuring retinal stray light has been developed. This test is based on the compensation comparison principle as explained in full earlier,[5] an enhancement of the direct compensation principle.[1] The test is intended for large scale routine clinical use. Therefore, assessment of the reliability of the test outcome is an important issue. This new test has been used in a European multicenter study (GLARE, see http://www.glare.be) to evaluate prevalence of visual impairment among 2422 drivers. Furthermore, this new technique has been used successfully to investigate the spectral nature of retinal stray light as function of age and pigmentation.[6]

It is the purpose of the present work to discuss the stochastic properties of a compensation comparison stray-light mea-

Address all correspondence to Joris Coppens, OST, NORI, Meibergdreef 47, Amsterdam, NH1105BA Netherlands. Tel: 3120 5665071; Fax: 3120 5666121; E-mail: j.coppens@ioi.knaw.nl

1083-3668/2006/11(3)/034027/9/$22.00 © 2006 SPIE

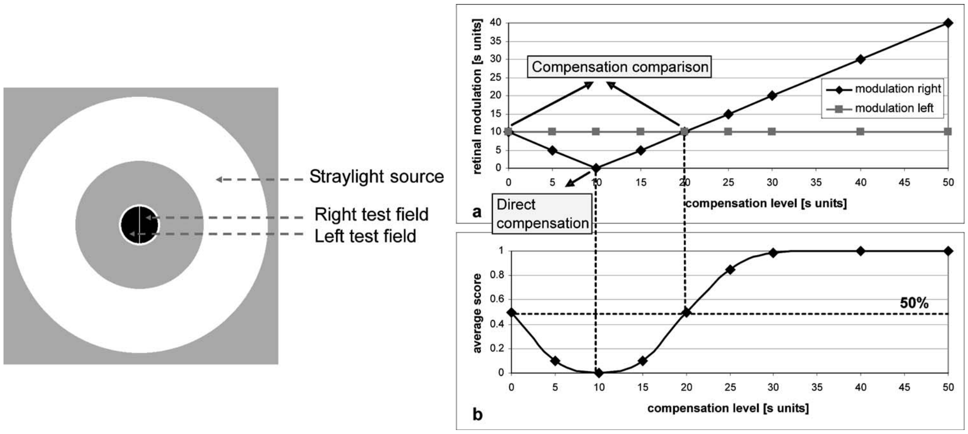

Fig. 1 Left side: field of view in a compensation comparison stray-light meter. The subject is required to compare the two test fields with respect to observed flicker strength. Right side top: example of a test with compensation in the right test field, and no compensation in the left test field. On the *y* axis the retinal modulation depth is shown. This is given as an absolute value of light modulation, expressed in so-called s units, explained earlier.[5] In the left test field, retinal modulation equals the stray-light induced flicker. In the right test field, retinal modulation equals the sum of stray-light induced flicker and compensation flicker. Right side bottom: average of many binary responses as function of the compensation level. This curve is the psychometric function that describes the chance of obtaining a 1 response. The amount of stray light in this example is set to an arbitrary value of 10.

surement. Based on these properties, data analysis methods are discussed that were developed to optimize for maximum reliability of the test outcome. A parameter indicating measurement reliability was developed and validated.

1.1 *Compensation Comparison Measurement*

The compensation comparison method is explained in full elsewhere.[5] In short, the field of view for a compensation comparison stray-light measurement is shown in Fig. 1. An annulus-shaped stray-light source is presented flickering. Due to intraocular scatter, part of the light from this ring is not focused on its proper place on the retina, but spreads out to other areas on the retina, such as the center of the annulus. This center is the location of two test fields. The flickering ring induces a weak flicker in the test fields. In one of the test fields, counterphase flicker is added. This counterphase flicker can compensate the flicker induced by the stray-light source. The amount of counterphase flicker that has to be added to completely extinguish the flicker induced by intraocular light scatter directly gives the amount of stray light in an eye. This principle was used in the direct compensation method; in this method, an adjustment was made until the flicker perception in the test field disappeared.

In the compensation comparison method, counterphase compensation light is presented in only one of the test fields, not in the other one. The task for the subject is to indicate which of the two test fields flickers the strongest. The compensated half will be chosen when a large amount of compensation is presented (e.g., at test level 50 in Fig. 1); such a response is recorded as 1. The noncompensated test field half will be chosen when the compensation in the other half extinguishes the stray-light flicker of the eye being tested (in Fig. 1 at test level 10); such a response is recorded as 0.

A compensation comparison measurement consists of a series of trials at various levels of compensation. The first phase of the test consists of test levels separated by 0.1 log units. This phase is used to obtain an initial estimate of the stray-light value. Around this first estimate, a second test phase with 11 stimuli spaced by 0.05 log units concludes the test. The stray-light value is subsequently determined by a maximum likelihood estimate of all recorded responses. Examples of measurement results are given later in Sec. 2 Methods.

1.2 *Measurement Reliability*

An important aspect of a test of visual function is the reliability of the test outcome. Due to the stochastic nature of the responses, repeated measurements will not yield identical results. This depends, among others, on the number of trials in a test, but also on the observation ability of the subject. To investigate the reliability, repeated measurements of retinal stray light were obtained on a large number of subjects (see next). The compensation comparison method uses a two alternative forced choice (2AFC) measurement paradigm. In such a paradigm, one of two alternatives can be given as a response, recorded as either 0 or 1. These binary responses allow the use of well-known statistical techniques in psychophysics, such as maximum likelihood algorithms. The latter algorithms are based on a chosen psychometric function, and are therefore called parametric.[7] The psychometric function describes the chance of a 1 response, given a trial at a certain stray-light test level and the subject's (true) stray-light value.

For well-established visual tests, such as those measuring visual acuity and contrast sensitivity, the shape of the corresponding psychometric function is described abundantly in the literature.[8–11] Important to note here is that the aforementioned tests measure a threshold, i.e., the borderline of the stimulus level that can be seen. The new 2AFC compensation comparison stray-light measurement is quite different in this respect; it is a comparison of two stimuli, at least one of which is well above threshold level. Both stimuli are equally strong, near twice the true stray-light level of a subject (20 in Fig. 1), resulting in chance performance (50%). At the true stray-light level of that subject (10 in Fig. 1), one of the stimuli is zero, resulting in a near 0 value of the psychometric function. At even lower levels, the value of the psychometric function increases again. So, the psychometric function for this measurement has a more complicated shape when compared to that of the well-known threshold tests. Note that the 0 point corresponds to the value of the subject's stray light, and is a factor of 2 (or 0.3 log units) below the 50% point. As becomes clear later, attention will shift from the 0 point to the 50% point at twice the stray-light value in the new approach.

2 Methods

In the present work, patient data are described from the multicenter GLARE study (see http://www.glare.be). Five centers participated, spread over Europe (Department of Ophthalmology, Vrije Universiteit Medical Center Amsterdam; Universitair Ziekenhuis Antwerpen; Centro de Oftalmología Barraquer Barcelona; Universitätsklinik für Augenheilkunde und Optometrie Salzburg; and Universitäts-Augenklinik Tübingen). In total, 2422 subjects were included. With two tests per eye, and some missing values, a total of 9340 stray-light measurements resulted and are used in this work. The measured population consisted of a wide range of subjects, including ages from 20 to 85, visual acuities from below 0.5 (logMAR 0.3) to more than 1.0 (logMAR 0.0), visual field defects, and ocular pathologies such as glaucoma and cataract. The total (2422 subjects) dataset has a wide range of differences in repeated stray-light values, making it very a suitable as a study object for measurement reliability. Stray-light measurements were performed as part of this study into the prevalence of visual impairment in drivers. The overall results of the GLARE study are reported separately. The study adhered to the guidelines of the Declaration of Helsinki for research in human subjects.

Given the 0 and 1 responses to a set of trials at various test levels, two questions have to be answered. 1. What is the best estimate of the true stray-light value of the subject tested? 2. How reliable is this estimate? Both questions can be answered by a likelihood analysis. To explain the principle of such an analysis, a very simple (fictive) dataset is used, containing only seven responses, at equidistant test levels, shown in Fig. 2 as filled circles. A psychometric function with arbitrary (assumed) transition level of the subject is shown in the upper left part of this graph with a continuous line, centered at test level 0.4, denoted with a vertical dotted line. In this simple example, an arbitrary shape for the psychometric function, was used. For the plotted psychometric function, we can calculate the chance of obtaining the seven responses shown. For each response, the chance is indicated with a vertical bar. For the 1 responses, this is the distance from 0 to the value of the psychometric function at the respective test level. For the 0 responses, this is the distance from 1 to the value of the psychometric function at the respective test level. Since the psychometric function describes the chance of a 1 response, the chance of a 0 response equals 1 minus the chance of a 1 response. The likelihood of obtaining the seven shown responses, assuming the shown psychometric function, is the multiplication of the chances for each response. This likelihood is shown in the lower half of the upper left quadrant of Fig. 2.

The assumed transition level of the psychometric function in the prior example was arbitrarily chosen. Maybe our subject had a different transition level than the one assumed. Looking at the upper right quadrant of Fig. 2, transition level 1.0 shows a higher likelihood than the lower transition levels. The lower left quadrant of Fig. 2 shows that for very high transition levels (e.g., 1.7), the likelihood is very low. Using a denser sampling, the maximum of the likelihood function is obtained at 1.35, shown in the lower right quadrant of Fig. 2.

Note that the 0 response at test level 1.0 seems to be a false response (outlier). In the psychometric function used in this example, a 5% rate for this kind of mistakes is assumed. The 0 response at test level 1.0 has virtually no influence on the location of the maximum, so outliers do not influence the result. This example suggests that the maximum of the likelihood function may be a robust estimate of the true transition level.

As explained in the Introduction in Sec. 1, an estimate of the transition level should be accompanied by a measure of its reliability. Several reliability measures were formulated and tried on the dataset from the GLARE study. Some of these measures were independent on knowledge of the psychometric function (nonparametric measures). Although nonparametric methods may be preferred in psychophysics, since with these methods no *a-priori* information is used, a parametric measure of reliability appeared to be most effective (see next).

As suggested by Harvey,[8] the width of the peak in the likelihood function can be used as a measure of reliability, and was used as a stopping criterion in his adaptive procedure ML-PEST. Asymptotically, that is, for a large number of trials, the shape of the likelihood function will approach that of a Gaussian.[12] For the relatively small number of trials in a compensation comparison measurement, the shape of the likelihood function may deviate from a Gaussian.

We determined the width of the likelihood function at four levels below the peak level, corresponding to confidence levels of 68, 95, 99.7, and 99.99%, respectively. Assuming a Gaussian shape for the peak of the likelihood function, these confidence levels correspond to a width of 2, 4, 6, and 8 standard deviations, respectively. Each width is divided by the number of standard deviations it represents, and then these four values are averaged. The resulting value is used as a reliability parameter called expected standard deviation (ESD).

The process of obtaining ESD is illustrated in Fig. 3 for a realistic set of stimuli. In the upper part of this figure, the responses are shown with a dot for the data from the initial phase, and with a cross for data from the final phase of the measurement. In this figure, a more realistic psychometric function is introduced.[5] The continuous line is the psychometric function at its most likely horizontal position.

In the lower part of Fig. 3, the likelihood function is shown. Note that the vertical scale is logarithmic. On a logarithmic scale, a Gaussian resembles a parabola. Furthermore, the likelihood has been normalized such that the maximum is 1. The horizontal bars indicate the width of the likelihood function at the four sample levels. The maximum of the likelihood function indicates the best estimate of the stray-light value. Note that this value lays a factor of 2 (or 0.3 log units) below the 50% point of the psychometric function.

In Fig. 4 (left), an example is given of a fair measurement, showing a region with overlapping 0 and 1 responses. Figure 4 (right) shows an example of a bad measurement, with re-

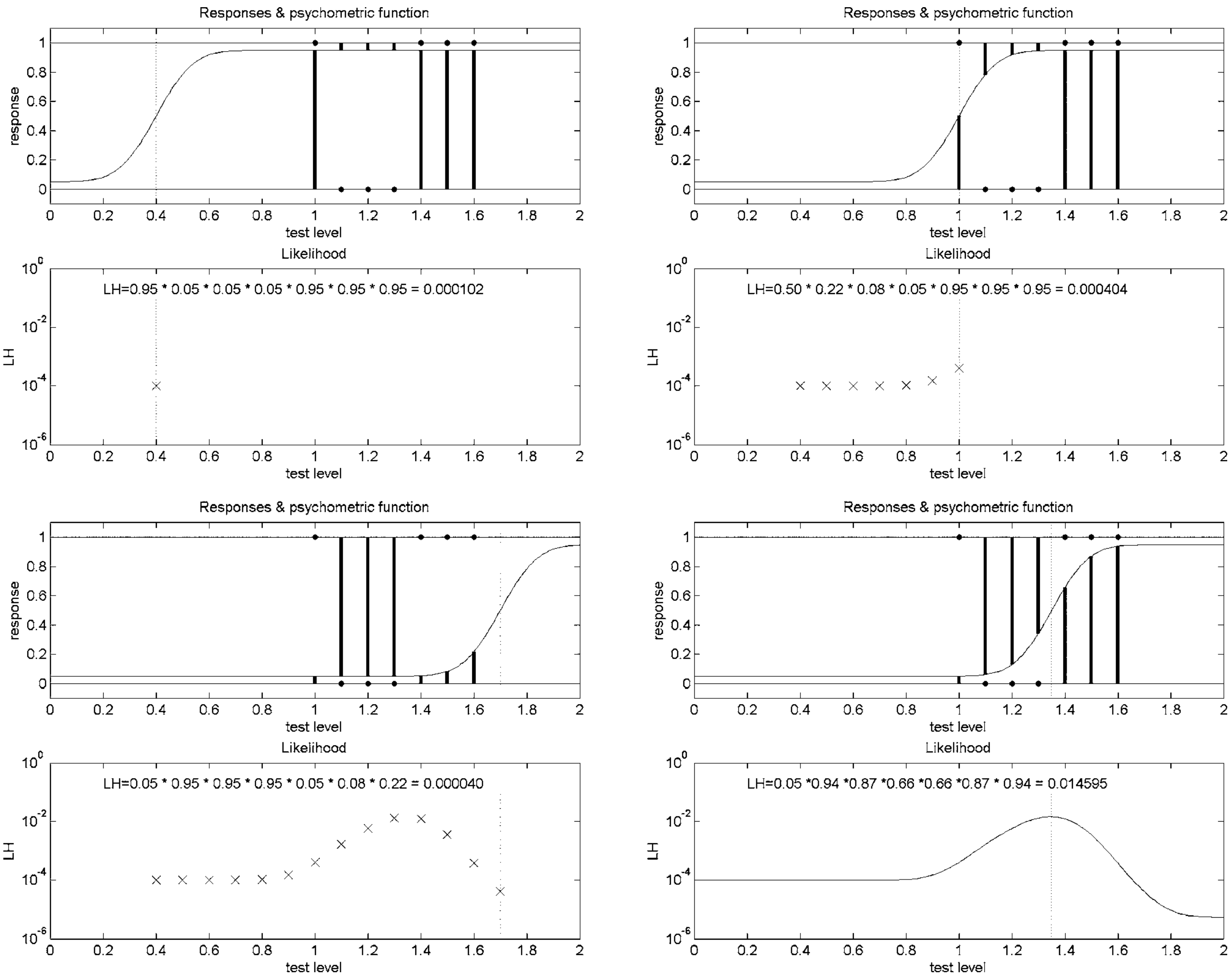

Fig. 2 Simplified example of a maximum likelihood fit. Top left: the upper half of the graph shows seven responses from a hypothetical 2AFC test, with filled circles. The continuous S-shaped curve indicates an (arbitrarily shaped) psychometric function for this test. Its horizontal position was chosen arbitrarily at a transition level of 0.4. The chance for each obtained response is indicated with the length of the bars. The likelihood for all seven responses is the multiplication of the chances for the individual responses. This likelihood is plotted with an X in the lower half of the plot. Top right: if the psychometric function is moved to a transition level of 1, the likelihood starts to increase. Bottom left: at transition level 1.7, the likelihood is beyond its maximum near 1.3. Bottom right: using a denser sampling, a maximum is found at transition level 1.35.

sponses that do not yield a reliable estimate of the stray-light value. Correspondingly, the peak of the likelihood function is ill defined in this case. At the lower two confidence levels of 99.7 and 99.99%, no valid estimate of the width can be obtained. The resulting ESD of 0.42 is an artificial value, and well above a reasonable maximum of 0.15.

This concludes the explanation of how measurement reliability is obtained from the 0 and 1 responses in a compensation comparison stray-light measurement. Verification of the value of ESD as a reliability measure was done with population data from the GLARE study. The result of the verification of ESD is given here in Sec. 3. The total (2422 subjects) dataset has a wide range of differences in repeated stray-light values, making it very suitable as study object for measurement reliability. During the course of the study, two different versions of the stimulus presentation were used. The instruction stimuli and the flicker levels presented in the initial phase were improved. 1073 subjects were tested with this newer version.

Figures 3 and 4 show an *a-priori* chosen shape for the psychometric function, based on the results of trial experiments.[5] Other shapes for the psychometric function are considered later. The shape of the psychometric function is independent of the stray-light level when logarithmic scales are used.[5] Only the horizontal position of the psychometric function is different for different stray-light levels. The shape itself is determined with two parameters: critical modulation depth contrast (MDCc) and delta.[5] MDCc is a parameter determining the steepness of the psychometric function. The steepness is proportional to the reciprocal of MDCc. Delta is a parameter describing the lapse rate; the percentage of accidental mistakes. Parameters of the *a-priori* shape of the psychometric function are: MDCc=0.16 and delta=0.05.

Repeated measure standard deviation was calculated in the usual way. The difference between two repeated measurements was determined for each eye, and then the standard deviation of this series of differences was calculated. Finally,

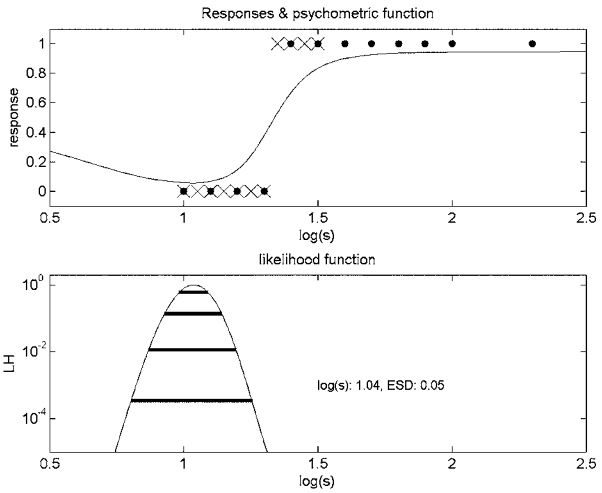

Fig. 3 Example of a perfect compensation comparison measurement. Responses are shown in the upper half with filled circles for the initial phase, and crosses for the final (refinement) phase of the test. The psychometric function is the S-shaped continuous line, plotted at the most likely horizontal position. The likelihood (ratio) function is plotted in the lower half of the figure, with horizontal bars indicating the four confidence levels used for calculation of expected standard deviation (ESD).

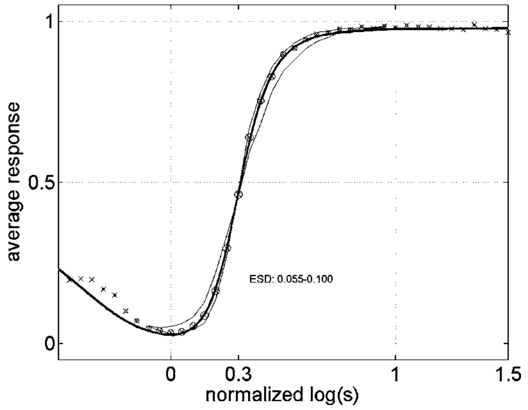

Fig. 5 Experimental psychometric function based on all measurements with ESD<0.10. The binary responses of these measurements were shifted according to each eye's stray-light value [log(s)]. Then, these responses were averaged in 0.05 log unit wide bins. The result is shown with crosses. Due to the measurement strategy, most responses were collected near the transition from 1 to 0. The amount of responses (weight) is indicated by the area of the circle at each point. The thick line is a maximum likelihood fit of a model function to the normalized binary responses. The two thin lines indicate two extreme possibilities for the dataset, which follow from the alternative normalization strategies explained in the text.

this standard deviation was divided by $\sqrt{2}$ to account for the fact that it originates from two independent measurements.

3 Results

3.1 *Experimental Psychometric Function*

Figure 5 shows the psychometric function, averaged over a large part (worst data excluded, see next) of the 2422 subjects. Before averaging, data were shifted along the log(s) axis, to normalize the data for differences in the stray-light value of the individual eyes. Originally, we normalized on the log(s) of the respective measurement itself, but then realized that this could give some bias. The maximum likelihood fit to obtain this log(s) value might act as a kind of matched filtering, resulting in a too steep estimate of the psychometric function. To prevent this, the (independent) fellow measurement outcome of each eye was used for normalization. However, this could also give bias, but in the opposite direction. In this case, due to the finite measurement accuracy, the measurement "jitter" will result in a too shallow estimate of the psychometric function. Finally, the average log(s) of the two repeated measurements was used for normalization. So, we used three alternative ways of normalization: 1. based on log(s) from the measurement itself; 2. based on log(s) from the fellow measurement; or 3. the average between these two log(s) values. The results for all three alternatives are shown in Fig. 5, with

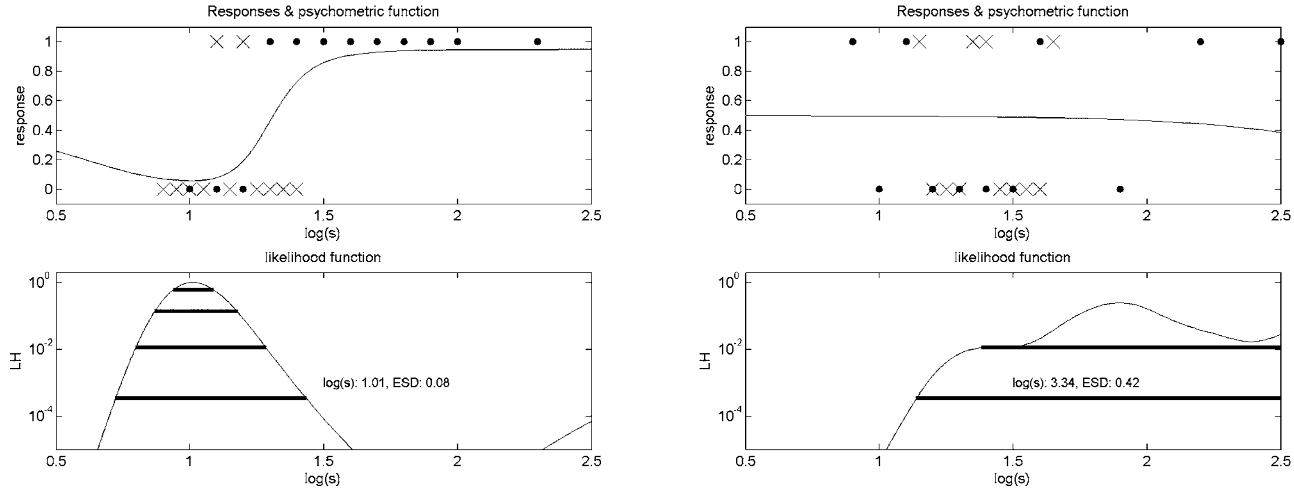

Fig. 4 Left side: example of a fair measurement. There is a region where 0 and 1 responses overlap. Right side: example of a bad measurement. In this case, no reliable estimate of the stray-light value can be found. The peak of the likelihood function is very wide, and the lower two confidence levels for the ESD are not bounded by the likelihood function.

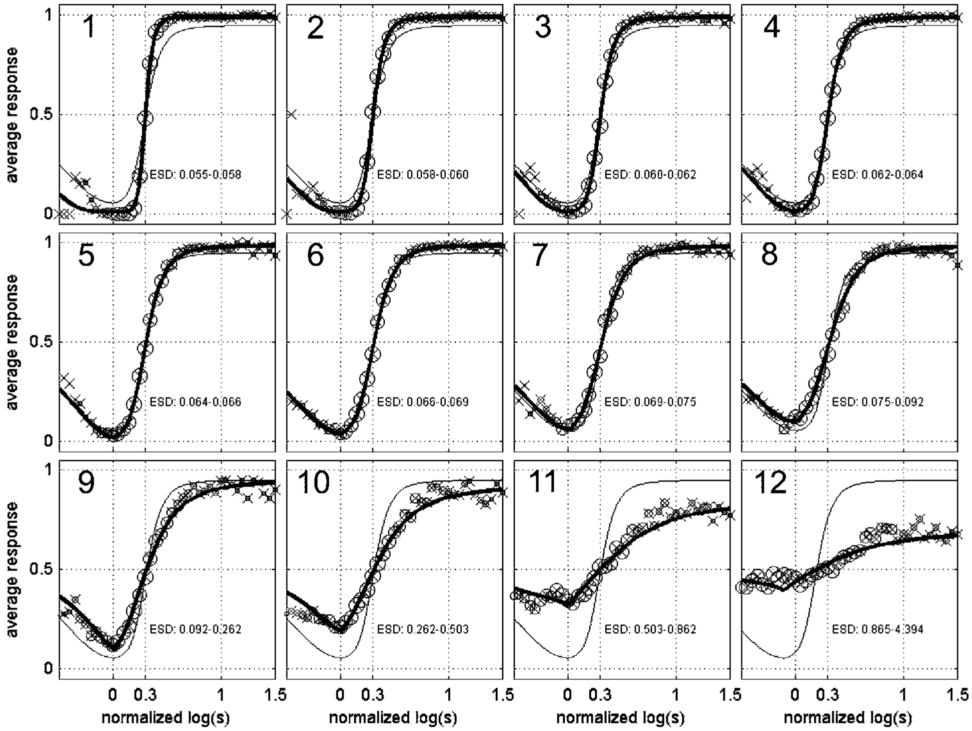

Fig. 6 Experimental psychometric function for various ranges of ESD. Data have been sorted according to ESD, and split into 12 equally sized groups. With increasing ESD, the psychometric function becomes less steep. The thick line is a fit to the data. The thin line is the (fixed) *a-priori* shape of the psychometric function that was used for obtaining log(s) and ESD values.

symbols for the alternative 3, and with continuous thin lines for the alternatives 1 and 2. Clearly, the differences are not large, and alternative 3 was used for further analysis. The shape of this psychometric function lies between those for the two other alternatives.

Figure 5 shows the result, sorted into 0.05 log unit wide bins and averaged. The crosses show the data points normalized on the mean log(s) of the two measurements. The area of the circles indicates the number of responses (weight). The largest circles are averages of more than 10,000 responses. The two thin lines show the result for the other normalization strategies, and may serve as limits for the true shape. The thick line is a maximum likelihood fit of the model for the psychometric function.[5]

In Fig. 5, all data have been used with an ESD of 0.10 or lower. These were considered sufficiently reliable (see next). It must be realized that in reality, different individuals may have different psychometric functions. To study potential differences in the shape of the psychometric function, the data were sorted according to ESD, and divided into 12 subsets, with an equal number of measured eyes (389) in each subset. The resulting shapes of the psychometric function are shown in Fig. 6. The thick line is a fit of the model function.[5] The thin line is the fixed shape used to obtain log(s) and ESD. These data may suggest that the shape of the psychometric function is not the same for the different subgroups.

3.2 *Optimum Psychometric Parameters*

Results shown earlier were based on an *a-priori* choice for the shape of the psychometric function. Perhaps, the results found can suggest improvements that will give better results than the *a-priori* choice. The assumption of constant shape of the psychometric function is not rigidly valid, as suggested by Fig. 6. Let us accept for the moment that a fixed psychometric function is adopted for data analysis. This raises the question what shape of the psychometric function should be chosen to yield optimum results for the population as a whole. In other words: what function gives the smallest repeated measure standard deviation?

To study this question, measurement pairs were sorted according to the maximum of the two ESD values of the (repeated) measurements. In Fig. 7, repeated measure standard deviations are plotted, starting from lowest ESD at the right, and including more and more of the measurements with higher ESD values (cumulative standard deviation). At the extreme left of Fig. 7, all data are included, also the worst.

In Fig. 7, the cumulative standard deviation is shown, as obtained with four different shapes of the psychometric function. The shapes of the psychometric functions are those from the first, fifth, and eighth dataset in Fig. 6, and the *a-priori* shape ((MDCc=0.16, delta=0.05. Fitted parameters of the experimental psychometric function are (from steep to shallow) MDCc=0.08, 0.16, 0.28, and delta=0.02, 0.02, and 0.08.

For a reliability criterion of 0.1 log units, the different choices of psychometric functions show only very subtle differences in the fraction of data that have to be excluded. This fraction (about 17%) is given by the horizontal position where the cumulative standard deviation crosses the $y=0.1$ line.

The similarity of the results obtained with the different shapes of the psychometric functions came as a surprise to us. Given this similarity, there was no reason to abandon the

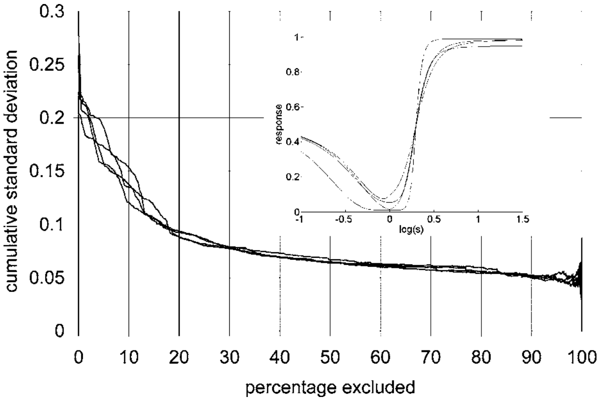

Fig. 7 Cumulative standard deviation for three experimental shapes of the psychometric function, and the *a-priori* shape. On the right extreme end, only the best measurements are included. Starting at this end, going left, more and more measurements are included, until at the extreme left end all measurements are included in the calculated standard deviation. The inset shows the shapes of the psychometric functions that gave these results.

a-priori shape in the further analyses.

In practical applications, ESD can be used to filter out the unreliable results. We applied a limit value of 0.1 in the GLARE study. The effect is illustrated in Fig. 8. The measurement pairs considered reliable are shown in Fig. 8 (left side), with the stray-light value of the first measurement along the x axis, and the stray-light value of the second measurement along the y axis. Two ellipses have been added that summarize the data; they indicate the 68 and 95% confidence limits of the data. For an ideal measurement accuracy, all data would lie on the $y=x$ line, and correspondingly, the ellipses would have zero width. For real measurements, the ratio of width and length of the ellipses indicates correlation of the two (repeated) measurements. The right side plot in Fig. 8 is comparable to the left side, but now for the fraction of data considered unreliable. Comparison of the correlation coefficients in the left ($r=0.79$) and right ($r=0.05$) side of Fig. 8 indicates that ESD is an effective filter for inclusion of reliable data.

3.3 *Expected Clinical Performance*

Demands for clinical use of a test may be stricter than those for a population study; results must be reliable for an individual patient as opposed to the average of a population. Therefore, the cumulative standard deviation as shown in Fig. 7 has limited clinical relevance. More important is how ESD is related to the measurement uncertainty of an individual patient. Again, the repeated measurements from the GLARE study were used. Only now the analysis will be restricted to the improved final version of the method (see Methods in Sec. 2), to better reflect the performance to be expected in future clinical use of the test.

The *a-priori* psychometric function was used to obtain log(s) and ESD values. The differences of the two log(s) values of repeated measurements were sorted according to the maximum of the two ESD values, as before. But now, repeated measure standard deviation was calculated over a window of 100 measurement pairs. This window was shifted from lowest ESD to highest ESD, like a moving average. Results are shown in Fig. 9. Note that in this figure 2049 measurement pairs are included, so a 100 point average corresponds to 5% along the horizontal axis. Assuming a clinically relevant limit value for the standard deviation of 0.1, Fig. 9 shows that in 13% of the eyes, at least one of the two measurements was substandard. This value follows from the percentile where the repeated measure standard deviation crosses the $y=0.1$ line.

3.4 *Relation Expected Standard Deviation and Repeated Measure Standard Deviation*

Figure 10 shows the repeated measure standard deviation that was also given in Fig. 9. In Fig. 10, however, it is plotted as a function of ESD. Note that the density of points is high for the lower ESD values (0.05 to 0.07), and much lower for ESD

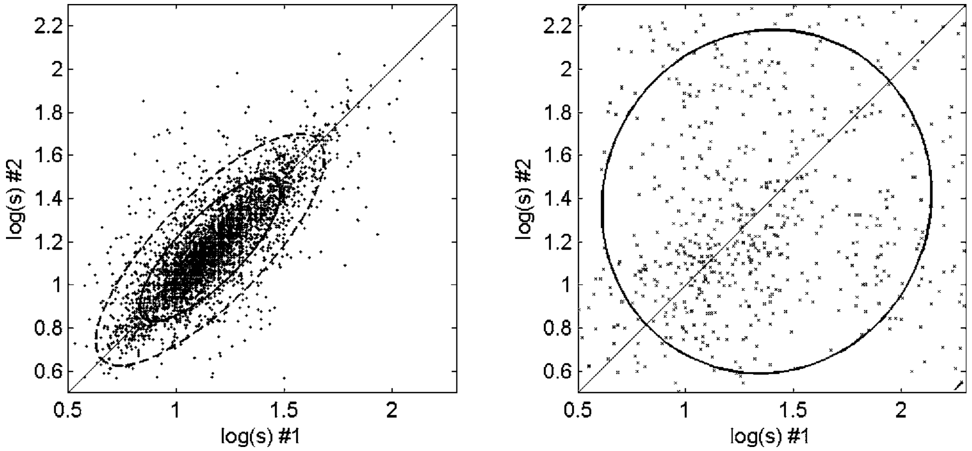

Fig. 8 Left side: scatter plot of repeated stray-light measurements; the stray-light value from the first measurement is plotted against that of the second measurement. Only data where both measurements were considered reliable are plotted. The correlation between the two measurements is summarized by the ellipses. The continuous ellipse indicates the 68% confidence limit, and the dashed ellipse the 95% confidence limit. Right side: similar plot, but now for the data considered unreliable.

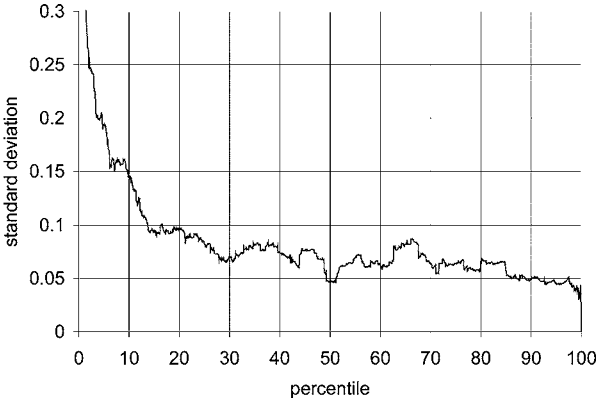

Fig. 9 Repeated measure standard deviation sorted according to ESD. Repeated measure standard deviation was calculated in a window of 100 measurements. The result is kind of a moving average over the true standard deviation, sorted according to ESD. Highest ESD results (worst performance to be expected) are on the left side; lowest ESD results (best performance to be expected) are on the right side.

>0.07. The points coarsely follow the continuous $y=x$ line that represents identity.

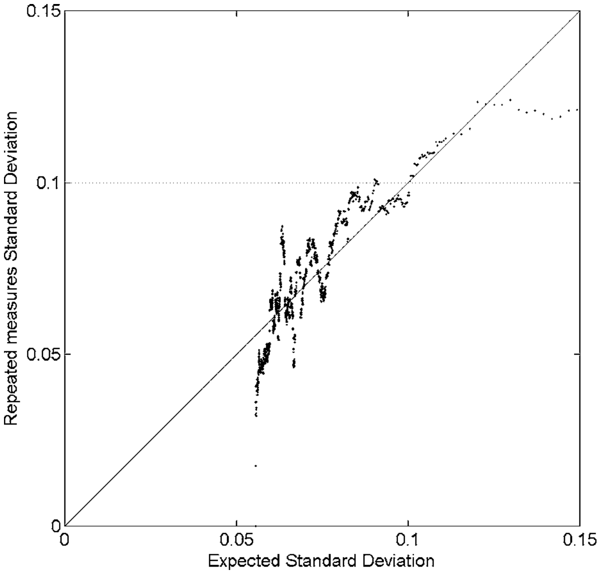

Fig. 10 Repeated measure standard deviation plotted as function of ESD. The dataset is the same as that in Fig. 9. The continuous $y=x$ line represents identity of true standard deviation and ESD. The dotted horizontal line represents a limit of 0.1 log unit standard deviation.

4 Discussion

For a psychophysical test, a measure indicating reliability of the test result is desirable; after all, the test result depends, for the major part, on accurate observations of the subject. This reflects both the power and the weakness of a psychophysical test. On the one hand, the result directly represents in a quantitative way a functional aspect of a subject's vision. On the other hand, reliability of the test result depends on human factors, such as explanation of the task, experience with the task, and mental state of the subject (is the subject paying attention to the task?), etc.

Ideally, psychophysical measurements are done by experienced observers in a laboratory environment. For this ideal situation, usually one shape of the psychometric function is assumed for all observers, and may be a near valid assumption. For a population study, it is obvious that there are differences in observation ability between individuals, and differences in the circumstances under which the measurements were done.

Such differences in observation ability have been found in this study. Data were divided in groups, sorted according to ESD. A maximum likelihood fit of a model function for the psychometric function for a compensation comparison straylight measurement shows that poor observers might have a higher threshold for flicker discrimination, as well as a higher lapse rate.

Given the found differences of the psychometric function between good and poor observers, and the central role that this function plays in the maximum likelihood analysis, one might be tempted to adapt the shape of the psychometric function in each individual test. Instead of having only the straylight value as a free parameter in the likelihood estimation, also the steepness and lapse rate of the psychometric function could be free parameters. This approach was tried, but abandoned in an early stage of the study. The number of trials in a test is limited to a practical value of 25. This number of samples appeared to be insufficient for estimation of more than one degree of freedom in the maximum likelihood analysis. So, only the stray-light level should be determined. When more degrees of freedom were allowed, the accuracy of the estimation of the straylight level would decrease unacceptably. Luckily, the results showed no need to allow more degrees of freedom.

Thus, calculation of ESD is based on a single assumed shape of the psychometric function. In accordance with the theory laid out in the literature,[8,12] this is (asymptotically) correct, assuming the shape of the psychometric function to be known. However, there are very good and very bad observers, with a corresponding change of shape of the (observer dependent) psychometric function. This raises the question how correct the value obtained for ESD is, since it is based on a single assumed psychometric function for all observers. Presently, we cannot offer a good answer to this question. To judge data reliability, ESD has proven to be of great value, as can be seen in Fig. 8. Its precise meaning as predictor for the true standard deviation to be expected from an observer is the subject of further study. For practical purposes, it was important to find that different shapes of the psychometric function turned out to have surprisingly little effect on the repeatability of the analysis outcome, and on the effectiveness of ESD as reliability criterion.

During the GLARE study, the stimulus design was slightly modified. A stronger flicker was presented in the initial phase of the test. A stronger flicker is more clearly perceived, important especially for the group of poor observers. Apart from stimulus design, feedback to the operator was also improved, such that the responses of the subject could be monitored. In the case of erratic responses from the subject, the comparison task could be re-explained. However, during data collection in the GLARE study, ESD was not available yet as criterion to redo a measurement. Interpretation of the "raw" binary responses appeared to be difficult for the operators in the field, resulting in repetition of the measurement only in extreme cases. This difficulty of interpretation of test responses as found in the field emphasizes the importance of having a number indicating measurement reliability.

The use of ESD in clinical practice may be to check whether the subject understood the task. Assuming a clinically relevant limit value for the standard deviation of 0.1, Fig. 9 shows that in 13% of the eyes at least one of the two measurements was substandard. This value follows from the percentile where the repeated measure standard deviation crosses the $y=0.1$ line. In practice, one would use the individual ESD of the measurement. Using a limit value of 0.1 for ESD, 9.8% of the measurements should have been redone in the GLARE study. With further improvements since the GLARE study, this value may drop.

The compensation comparison method for measuring retinal stray light and the maximum likelihood analysis described in this work have been implemented by Oculus GmbH in a commercially available instrument, called C-Quant. Instead of the cathode ray tube (CRT) that was used for stimulus presentation in the GLARE study, dedicated hardware was developed. This dedicated hardware allows presentation of better defined stimuli. Most notably, the intrinsic stray light of the C-Quant is negligible, and the luminance is a factor of 3 higher than in the CRT implementation. Preliminary evaluation of this device has shown an improved rate of reliable results. In this instrument, the ESD limit value has been set to 0.08.

In conclusion, the binary responses obtained in a compensation comparison stray-light measurement can successfully be used for an accurate estimate of the stray-light value, as well as a measure of reliability of this stray-light value. Analysis is based on a single chosen shape of the psychometric function. Although the population data suggest a wide range of shapes for the psychometric function, using these various shapes is unnecessary, as the likelihood analysis gives similar results.

Acknowledgments

This study was supported by grant SUB-B27020B-E3-GLARE-2002-S07.18091 of the European Commission.

References

1. T. J. T. P. van den Berg, “Importance of pathological intraocular light scatter for visual disability,” *Doc. Ophthalmol.* **61**, 327–333 (1986).
2. J. J. Vos, “Disability glare—a state of the art report,” *Commission Intl. l’Eclairage J.* **3/2**, 39–53 (1984).
3. T. J. T. P. van den Berg, B. S. Hwan, and J. W. Delleman, “The intraocular straylight function in some hereditary corneal dystrophies,” *Doc. Ophthalmol.* **85**, 13–19 (1993).
4. T. J. T. P. van den Berg, “Analysis of intraocular straylight, especially in relation to age,” *Optom. Vision Sci.* **72**, 52–59 (1995).
5. L. Franssen, J. E. Coppens, and T. J. T. P. van den Berg, “Compensation comparison method for assessment of retinal straylight,” *Invest. Ophthalmol. Visual Sci.* **47**, 768–776 (2006).
6. J. E. Coppens, L. Franssen, and T. J. T. P. van den Berg, “Wavelength dependence of retinal straylight,” *Exp. Eye Res.* **82**, 688–692 (2006).
7. B. Treutwein, “Adaptive psychophysical procedures,” *Vision Res.* **35**, 2503–2522 (1995).
8. L. O. Harvey, Jr., “Efficient estimation of sensory thresholds,” *Behav. Res. Methods Instrum. Comput.* **18**, 623–632 (1986).
9. G. Westheimer, “Scaling of visual acuity measurements,” *Arch. Ophthalmol. (Chicago)* **97**, 327–330 (1979).
10. K. R. Alexander, W. Xie, and D. J. Derlacki, “Visual acuity and contrast sensitivity for individual Sloan letters,” *Vision Res.* **37**, 813–819 (1997).
11. J. Nachmias, “On the psychometric function for contrast detection,” *Vision Res.* **21**, 215–223 (1981).
12. W. Q. Meeker and L. A. Escobar, “Teaching about Approximate Confidence Regions Based on Maximum Likelihood Estimation,” *Am. Stat.* **49**, 48–53 (1995).