Reliability of the compensation comparison method for measuring retinal stray light studied using Monte-Carlo simulations

Joris E. Coppens
Luuk Franssen
Thomas J. T. P. van den Berg
Netherlands Ophthalmic Research Institute
Meibergdreef 47 1105BA
Amsterdam, Netherlands

Abstract. Recently the psychophysical compensation comparison method was developed for routine measurement of retinal stray light. The subject's responses to a series of two-alternative-forced-choice trials are analyzed using a maximum-likelihood (ML) approach assuming some fixed shape for the psychometric function (PF). This study evaluates the reliability of the method using Monte-Carlo simulations. Various sampling strategies were investigated, including the two-phase sampling strategy that is used in a commercially available instrument. Results are given for the effective dynamic range and measurement accuracy. The effect of a mismatch of the shape of the PF of an observer and the fixed shape used in the ML analysis was analyzed. Main outcomes are that the two-phase sampling scheme gives good precision (Standard deviation=0.07 logarithmic units on average) for estimation of the stray light value. Bias is virtually zero. Furthermore, a reliability index was derived from the responses and found to be effective. [DOI: 10.1117/1.2357731]

Keywords: glare; straylight; point spread function; psychophysics; maximum likelihood; eye.

Paper 06018R received Feb. 3, 2006; revised manuscript received Jun. 9, 2006; accepted for publication Jun. 13, 2006; published online Oct. 12, 2006.

1 Introduction

Recently a novel psychophysical method to measure retinal stray light was introduced. Details of this so-called compensation comparison (CC) method have been published earlier.[1] The CC method is a psychophysical approach to assess the amount of light scattered by the ocular media (e.g., cornea and crystalline lens) toward the retina, or to be more precise the stray light as it is sensed by the retina.[2] Retinal stray light is a disturbing effect to vision, resulting in complaints such as blinding by headlights while driving at night or hazy vision during day time.[3,4] The CC method can be used in clinical practice to determine the severity of pathological states, such as cataract and corneal edema, in a functional sense. The CC method has been implemented by Oculus GmbH in a commercially available instrument called C-Quant. It is the purpose of this paper to discuss the psychophysics involved in a CC test and to gain more insight in the stochastic behavior of the method.

The CC method works as follows: a subject is presented a stimulus as shown in Fig. 1. It consists of an annulus-shaped stray light source, and centered within this annulus there are two half-circular test fields. During a short trial period, the annulus flickers at 8 Hz. Due to intraocular light scatter, part of the light from the (strongly) flickering annulus is deflected, inducing a (weak) flicker in the two test fields. This deflected light is called stray light. The amount of stray light in an eye can be quantified by means of the (equivalent) luminance it induces in the test fields. More precisely, stray light is defined as the *equivalent luminance* normalized on the *illuminance* of the stray light source at the pupil plane.[4,5] So, when the induced flicker luminance in the test fields is known, the amount of stray light in an eye can be determined.

1.1 *Direct Compensation*

The induced flicker luminance in the test fields can be assessed by adding a *compensating* counterphase flicker luminance in the test fields. Originally, in the "direct compensation" method, this luminance was adjusted by the subject until the flicker perceived in the test fields was extinguished. The amount of counterphase luminance needed equals the equivalent luminance induced by the flickering source, giving a *direct* measure of the amount of stray light in an eye. This "direct compensation" method for measuring retinal straylight has been used as golden standard.[6] The "direct compensation" method, however, was not suitable for use in routine clinical practice.[7,8] Most notably, the direct compensation method lacked control over the adjustment strategy of the subjects, and no indication of the reliability of an individual adjustment result was available.

Address all correspondence to Joris Coppens, Netherlands Ophthalmic Research Institute, Meibergdreef 47 1105BA, Amsterdam, Netherlands; Tel: +3120-5665071; Fax: +3120-5666121; E-mail: j.coppens@ioi.knaw.nl

1083-3668/2006/11(5)/054010/12/$22.00 © 2006 SPIE

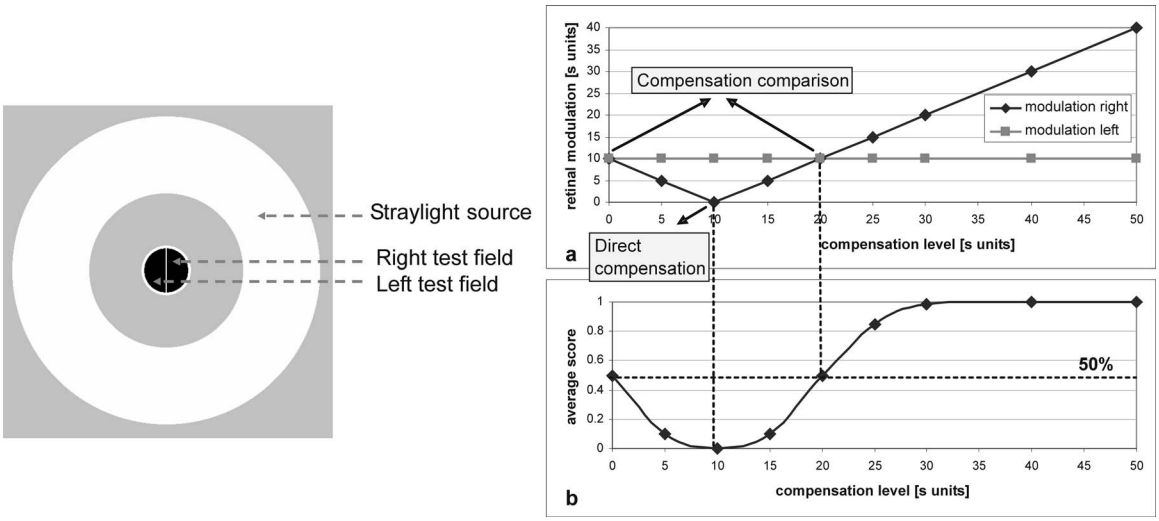

Fig. 1 (left) Stimulus layout presented in a CC test. The annulus shaped stray light source is presented flickering at 8 Hz during a trial. The source induces a (relatively weak) flicker in the central test fields, due to intraocular light scatter. (upper right) Retinal flicker modulation is plotted as function of compensation level. Assume that the right test field is given counterphase compensation flicker, and that the left test field is not compensated. (lower right) The PF that describes the average response as a function of compensation level. The task of the subject is to indicate which test field flickers strongest. If the compensated field is chosen, this is scored as 1. For high levels of compensation flicker (e.g., 50), the right field flickers clearly stronger than the left, resulting in a 1 response. When the stray light flicker is exactly compensated (at compensation level 10 in this example), there is no flicker in the right test field, resulting in a 0 response. When the compensation level is twice the stray light value (20 in this example) both test fields flicker equally strong, resulting in chance response (0.5).

1.2 *Compensation Comparison*

The CC method for measuring retinal stray light was developed to solve the problems met in clinical practice with the direct compensation method. The most important improvement is that the test follows a two-alternative-forced-choice (2AFC) paradigm. Instead of adjusting the compensating flicker luminance, a fixed number (25) of short duration trials (1 or 2 s) are presented. In these trials, only one of the test fields is given counterphase compensation flicker. So, in one of the test fields, only the induced stray light flicker is perceived, and in the other one, the combination of induced stray light flicker and added compensation flicker. The task for the subject is to compare the flickers perceived in both test fields and to indicate which of the two test fields flickers strongest. A choice in favor of the compensated field is recorded as 1, a choice for the uncompensated field as 0.

When the compensated test field is presented with a strong counterphase flicker, this field is chosen as flickering most, resulting in a 1 response (Fig. 1 lower right). When the counterphase flicker exactly compensates the induced stray light flicker, the perceived flicker in the compensated field is 0, and therefore, the *un*compensated field will be chosen as flickering most, resulting in a 0 response (at compensation level 10 in Fig. 1). When the compensated test field has *twice* the amount of induced stray light flicker (at compensation level 20 in Fig. 1), both test fields will have equal flicker strength. However, the subject is forced to give a response (0 or 1), and the *chance* of a 1 response will be 50%. The whole chance process is described by the psychometric function (PF). This function starts at 0.5 for no compensation, goes to (almost) 0 at exactly the compensation level, and rises to (almost) 1 for higher compensation levels (see Fig. 1 bottom right). In an earlier paper,[1] a mathematical formulation of the PF for the CC method is discussed; a summary of this formulation is given in the appendix of this report. The upper half of Fig. 2 shows the actual PF used and a set of responses obtained in a measurement.

1.2.1 *Sampling strategy*

In clinical practice, a relatively low number of trials in a test is desirable to minimize test duration. After some preliminary tests, we arrived at the following sampling strategy. The test starts with 12 initial trials. These trials are presented starting with a high level of compensation and subsequently have lower levels of compensation, spaced by 0.1 logarithmic units. So a subject will start responding with 1 and at lower compensation levels (lower than twice the stray light level of the eye tested) respond with 0. The transition from 1 to 0 responses is used to obtain an initial estimate of the stray light level. The test is then refined in a final phase, where 13 stimuli spaced by 0.05 logarithmic units are presented around the initially found transition level. These final trials are presented in random order. The range of initial trials can be set to seven levels (ranges A to G), depending on the stray light level expected (see Table 1). The ranges A to E follow the normal age dependence of stray light in healthy eyes.

1.2.2 *Maximum likelihood analysis*

A subject's stray light value in a CC test is determined using the binary 0 and 1 responses on basis of a maximum-likelihood (ML) analysis. In short, the likelihood of obtaining a 1 response is given by the value of the PF at the respective compensation level. The likelihood of obtaining a 0 response is given by 1-PF. The total likelihood of all 0 and 1 responses is then given by the product of the likelihood of all the single responses. One of the parameters of the PF is the stray light

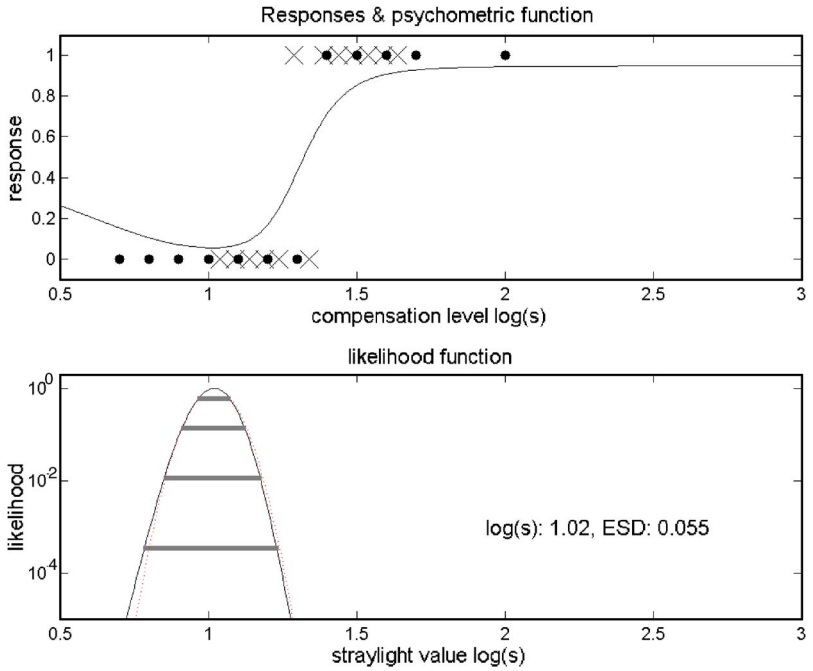

Fig. 2 Example of a CC stray light measurement with range setting A. The upper plot shows the raw 0 and 1 responses obtained as function of compensation level. The responses of the initial phase are shown as dots; the responses of the final phase are shown as crosses. The continuous line is the PF describing the chance of a 1 response. The PF is plotted at its most likely horizontal position for the responses shown. The lower plot shows the likelihood ratio function. The horizontal position of the maximum of this function indicates the most likely stray light value, given the responses shown in the upper plot. The thick horizontal lines show the levels where the width of the peak of the likelihood is determined for calculation of ESD. The dotted line is a Gaussian (resembling a parabola due to the logarithmic scaling of the *y* axis), with ESD as width parameter σ.

value. So, the total likelihood can be calculated as function of stray light value (see the bottom of Fig. 2). This likelihood function is normalized such that the maximum is 1. Such a normalized likelihood function is also known as a likelihood ratio function. A more elaborated explanation of the ML analysis is given in an earlier paper.[9] The stray light level corresponding to the top of the likelihood (ratio) function is used as the most likely estimate of the true stray light level.

Apart from estimation of the most likely stray light value, the likelihood function can be used to estimate the uncertainty of this value. We have called this the expected standard deviation (ESD). The calculation of this value is explained in more detail in Sec. 2. Here it may suffice to mention that the width of the peak of the likelihood function is evaluated at four levels below the maximum, shown by horizontal bars in the lower half of Fig. 2. The weighted average of these widths gives ESD.

ESD has proven to be useful to identify unreliable measurements during data analysis of the stray light measurements in the GLARE study.[9] Although a firm theoretical basis exists on likelihood ratio (as will be explained in Sec. 2), the initial development of ESD was heuristic. It must be noted here that the strict theory is based on assumptions about the PF and the sampling, both not necessarily valid in our application. However, in practice, ESD turned out to be the most effective criterion after evaluation of several different measures of reliability.

Table 1 Range settings for the stimuli presented in the initial phase of a CC measurement.

Range	Initial Compensation Levels Presented	Intended log(*s*) Range	Intended Use
A	2.0, 1.7, 1.6,...,0.7	≤1.1	Healthy eye (age ≤45)
B	2.1, 1.8, 1.7,...,0.8	0.8 to 1.2	Healthy eye (age 46 to 55)
C	2.2, 1.9, 1.8,...,0.9	0.9 to 1.3	Healthy eye (age 56 to 65)
D	2.3, 2.0, 1.9,...,1.0	1.0 to 1.4	Healthy eye (age 66 to 75)
E	2.5, 2.2, 2.1,...,1.2	1.2 to 1.6	Healthy eye (age ≥76)/early opacity
F	2.7, 2.4, 2.3,...,1.4	1.4 to 1.8	Moderate opacity
G	3.0, 2.7, 2.6,...,1.7	≥1.7	Severe cataract or corneal edema

1.2.3 *Individual dependent shape of PF*

As explained earlier, the PF has a central role in the estimation of both the stray light value and ESD. In the ML estimation, a single, fixed shape of the PF is used. However, analysis of the GLARE data suggests that the shape of the PF might be different between individuals. Figure 3 shows 12 experimen-

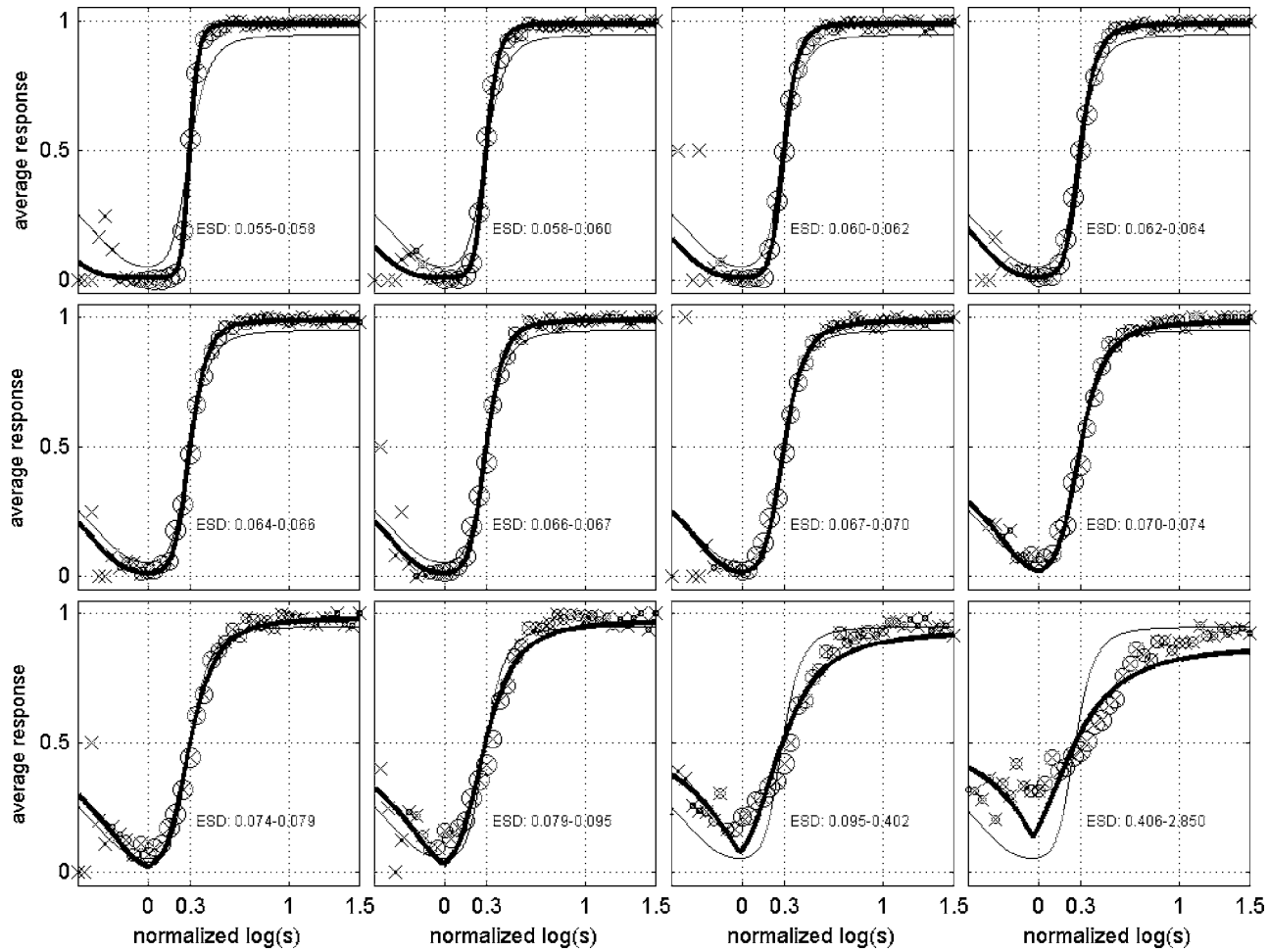

Fig. 3 Experimental PF obtained from 1073 subjects in the GLARE study. Data have been sorted according to ESD and split into 12 equally sized groups. The crosses indicate the average response. The number of responses (weight) is indicated by the area of the circles. The thick line is a ML fit of the PF model given in the appendix to the 0 and 1 responses. *MDCc* values obtained are (top row, from left to right) 0.070, 0.094, 0.114, 0.131; (middle row, from left to right) 0.139, 0.142, 0.170, 0.204; and (bottom row) 0.217, 0.259, 0.376, 0.521. The lapse rate λ was fixed at 0.01 during the fit. Before averaging and fitting the responses, the responses of each test were shifted along the horizontal axis, such that all responses were normalized to a stray light level log(*s*)=0. The stray light values of the individual eyes were determined with the PF shown as a thin continuous line (MDC_c=0.156, λ=0.05).

tal PF obtained from the GLARE study. Data from 1073 subjects have been sorted according to ESD and split into 12 equally sized groups. The responses in each group were binned and averaged after normalization on the individual stray light value of the eyes. These averaged responses were fitted with the PF described in the appendix (see also Refs. 1 and 9).

1.3 *Monte-Carlo Simulation*

Although the GLARE results have been a valuable source of information on the stochastic properties of a CC measurement, some relevant questions (listed below) cannot be answered *directly* with the GLARE data set. An essential shortcoming for answering these questions is that the *true* stray light value of the eyes tested is unknown. Therefore, possible systematic errors (bias) cannot be evaluated with these data. Furthermore, the *true* PF of the individuals in the study is unknown. Also, the effects of different sampling strategies could not be studied, nor the basics of the relationship between ESD and (true) SD.

The limitations mentioned above can be resolved by using Monte-Carlo simulations. Instead of analyzing responses from real subjects, responses are simulated by computer. In such simulations, both the assumed subject characteristics (stray light value and shape of PF) and the returned results are known. An additional benefit is that the input parameters (such as number and distribution of samples) can be varied as desired.

The purpose of this paper is to use Monte-Carlo simulations to study questions such as the following: (1) How well does ESD represent the true standard deviation? (2) What is the relation of standard deviation (SD) (and ESD) with number and spacing of the samples? (3) How effective is the two-phase sampling scheme described above? (4) Does the CC analysis introduce systematic deviations in the estimated stray light value? (5) What happens if there is a discrepancy between the PF of an observer and the assumed PF used in the

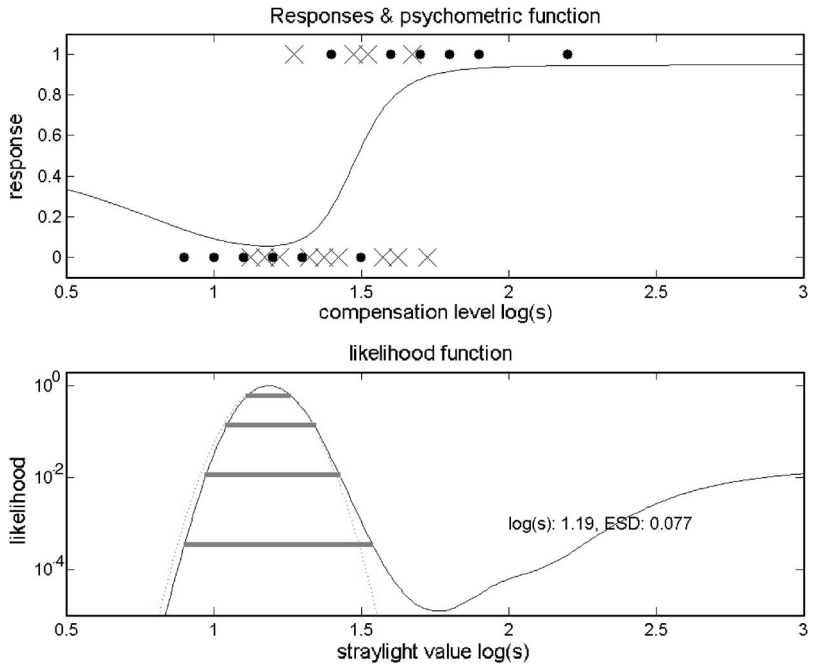

Fig. 4 Example of a CC stray light measurement, comparable to Fig. 2. The measurement range was C. In the upper plot, the responses of a poor observer are shown, as opposed to the responses of a good observer shown in Fig. 2. The continuous line is the PF, shown at its most likely horizontal position for the responses given. The lower plot shows the likelihood ratio function. When compared to Fig. 2, the peak is wider. The stray light value (log(s)=1.19) determined in this example has just acceptable expected accuracy (ESD=0.077).

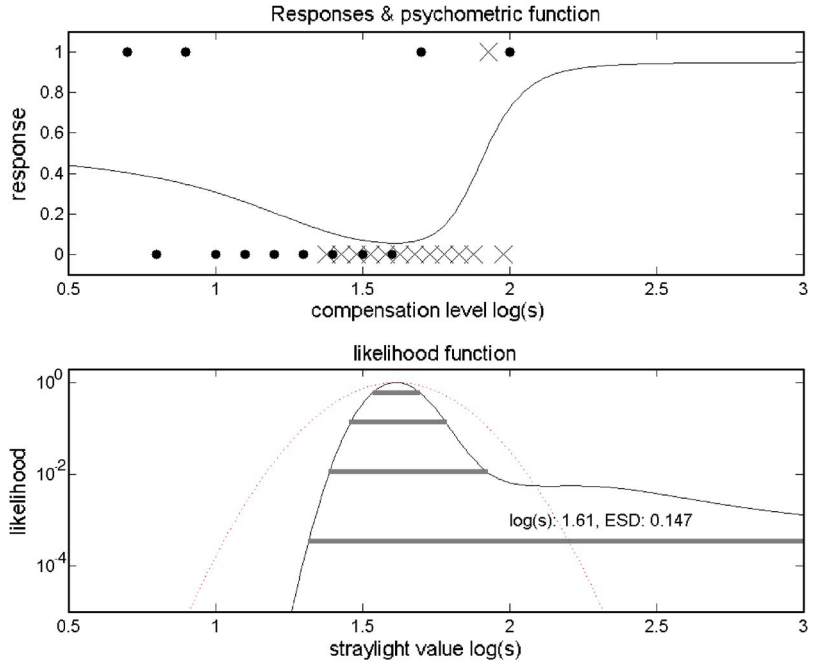

Fig. 5 Example of an unacceptable measurement, caused by a too low range setting (range A) for the initial phase of the measurement. With an estimated stray light value log(s)=1.61, the measurement should be redone in range F. Because of the erroneous range setting the samples of the final phase of the test are placed at too low compensation levels. As a result, the likelihood function does not bind the lowest likelihood level used for ESD calculation. The resulting ESD value is therefore very high. Also, the likelihood ratio function deviates largely from a Gaussian with width parameter σ equal to ESD, as shown by the dotted line.

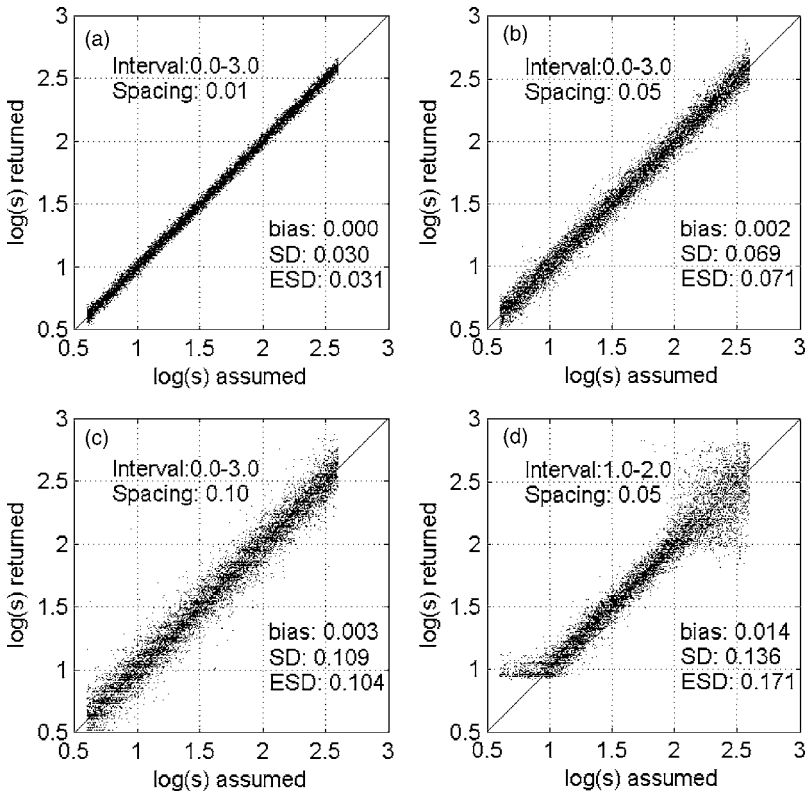

Fig. 6 (a) Monte-Carlo analysis of a CC measurement. On the *y* axis the result of the simulated measurement is shown, on the *x* axis the assumed stray light value. In the ideal case, both are identical and then lie on the $y=x$ line. A uniform distribution of 8000 stray light values ranging from $\log(s)=0.6$ to 2.6 is simulated with a range of trials at compensation levels $\log(s)$ from 0.3 to 3.3. The spacing of the trials is 0.01 logarithmic unit, resulting in 300 trials per test. (b) Similar to A, but now the spacing of the trials is 0.05 logarithmic units, resulting in 60 trials per test. (c) Similar to A and B, but now the spacing is 0.10 logarithmic units, resulting in 30 trials per test. Due to the rather coarse sampling, some discretization is seen in the simulated values. (d) The range of trials is now limited to compensation values from 1.3 to 2.3. Spacing is 0.05 logarithmic units, so there are 20 trials in a test. Stray light values outside the range of trials result in inaccurate estimates, as can be seen by the larger deviation from the $y=x$ line outside the tested range.

ML analysis? (6) What happens if there is a discrepancy between stray light value and sampling range? To answer these questions, three sets of Monte-Carlo simulations were generated, with increasing complexity of input parameters. Details of these simulations will be given in Sec. 2.

2 Methods

2.1 *ESD Calculation*

For a large number of trials, the shape of the likelihood ratio function will approach that of a Gaussian function.[10] This Gaussian function, when properly normalized (having an integrated value of 1), represents the *probability density function* of the most likely stray light value obtained.[11] For the relatively small number of trials in a CC test, the shape of the likelihood function may deviate from a Gaussian function. For an ESD calculation, the width of the peak of the likelihood ratio function is determined at four levels below the maximum value (normalized on a maximum value of 1). The levels used are 0.61, 0.14, 0.011, and 0.0003, respectively. At these levels, a Gaussian function with width parameter σ, has a (total) width of 2σ, 4σ, 6σ, and 8σ respectively. The corresponding confidence levels for these widths are 68, 95, 99.7, and 99.99%. ESD is calculated by averaging the four widths after dividing them by the number of SD they represent. Figures 2, 4, and 5 show examples of the widths found in real CC measurements. Note that the likelihood ratio functions in the lower plots of these figures have a logarithmic scale on the *y* axis. On a logarithmic scale, a Gaussian function resembles a parabola.

The example in Fig. 5 shows large deviations of the likelihood function from a Gaussian shape. In fact, the deviations are so large that the width of the peak determined at the lowest confidence level is not bounded by the likelihood ratio function. The corresponding ESD value is very large. The deviations in this example were caused by the use of an incorrect measurement range during the initial phase of the test, resulting in improper distribution of the trials.

2.2 *Monte-Carlo Simulation*

As already mentioned in Sec. 1, field tests of a psychophysical measurement method are not sufficient to fully analyze it. In Sec. 3, all trial responses have been generated by a computer "subject." Since the PF describes the *chance* of a 1 response for a subject, a computer "response" is easily generated with a uniform random number that is compared to the value of the PF. In total, three simulation settings are presented in this paper with increasing complexity and increasing relation with the real CC test. The first set of simulations presented was

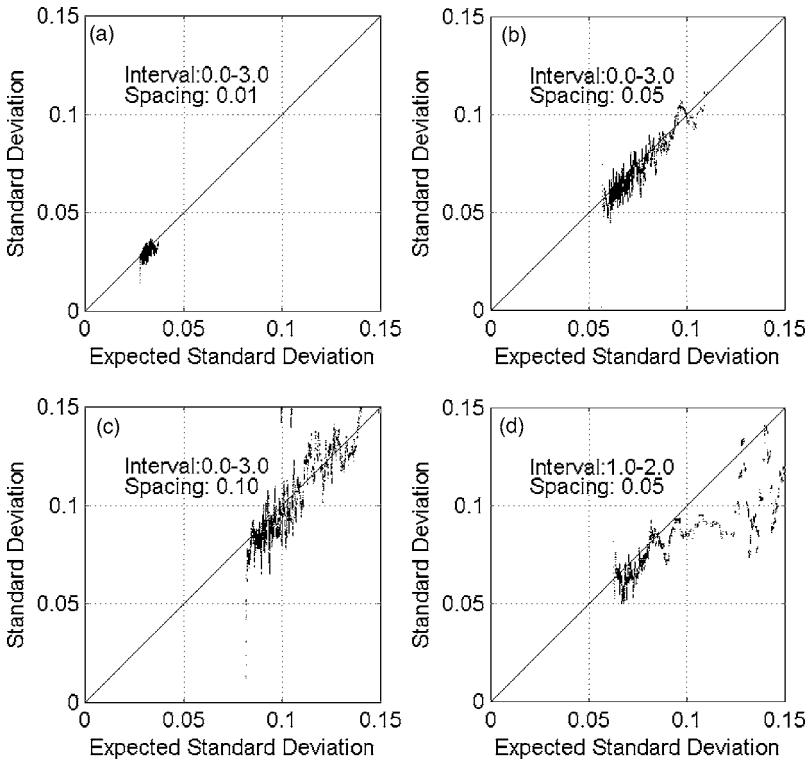

Fig. 7 Results from the same simulations as shown in Fig. 6. On the *y* axis, a moving average ($n=100$) of the SD of the difference between assumed and returned stray light value is shown. On the *x* axis, a moving average ($n=100$) of the ESD is shown, calculated with the likelihood function as explained in Sec. 2. In the ideal case, ESD represents the SD from the simulation, and all points would lie on the $y=x$ line.

created with an identical shape of the PF for generation of responses and ML analysis. Simulated straylight values in this set have a uniform distribution from $\log(s)=0.6$ to $\log(s)=2.6$. Also the compensation levels of the trials "presented" have a uniform distribution, mostly from $\log(s)=0.3$ to $\log(s)=3.3$. These simulations are used to investigate the effect of sampling density on measurement outcome. More concretely, the results of these simulations are used to show the influence of the number of trials on measurement accuracy (SD) and, furthermore, whether the resulting ESD is representative for the true SD or not. The second set of simulations is used to investigate the properties of the somewhat more complicated two-phase sampling scheme as described in Sec. 1. Special attention is given to the range settings from A to F that determine the compensation levels of the trials presented in the initial phase of a CC test. The third set of simulations is most complicated and intended to reproduce the results from the GLARE study. Simulated stray light values and range settings were taken from the GLARE data. In this simulation, the range was set according to age averages as given in Table 1. This last set of simulations was created with the 12 different shapes of the PF obtained from the GLARE data shown in Fig. 3. All data of the simulations presented were analyzed using the ML routines based on a single assumed (fixed) shape for the PF.[9]

3 Results

3.1 *Sampling Strategy*

Figures 6 and 7 show results for the ideal case when the PF assumed for the ML analysis is identical to the true PF of the (simulated) subject. Figure 6 shows the returned stray light value as function of the assumed stray light value in the simulation. Each simulation contains 8000 assumed stray light values, uniformly distributed from $\log(s)=0.6$ to $\log(s)=2.6$. Four different sampling schemes were used. The results presented in Figs. 6(a)–6(c) have uniformly distributed sampling, with compensation levels ranging from $\log(s)=0.3$ to $\log(s)=3.3$. For these results, the trial levels were spaced by 0.01, 0.05, and 0.1 logarithmic units, respectively, corresponding to a total number of trials per test of 300, 60, and 30. The results show how the accuracy of the test increases with the number of trials. The SD of the results is approximately proportional to the reciprocal of the square root of the sample spacing. Furthermore, the overall SD of the difference between assumed and returned stray light values closely follows the average ESD. The average difference of assumed and returned stray light values (bias) is not statistically significant. The fourth simulation [Fig. 6(d)] shows the result for an erroneous sampling scheme. The same spacing as in Fig. 6(b) was used, but the compensation levels range from $\log(s)=1.3$ to $\log(s)=2.3$, whereas the range of simulated stray light values was from $\log(s)=0.6$ to $\log(s)=2.6$, as before. Figure 6(d) shows how mismatch between stray light value and sample range upsets the estimate.

Figure 6 showed ESD to be equal to SD *on average*. But ESD (and SD) may differ between individual measurements. An important research question was whether ESD on an individual basis predicts SD. Figure 7 shows SD as function of ESD for the same simulations as shown in Figure 6. Both SD

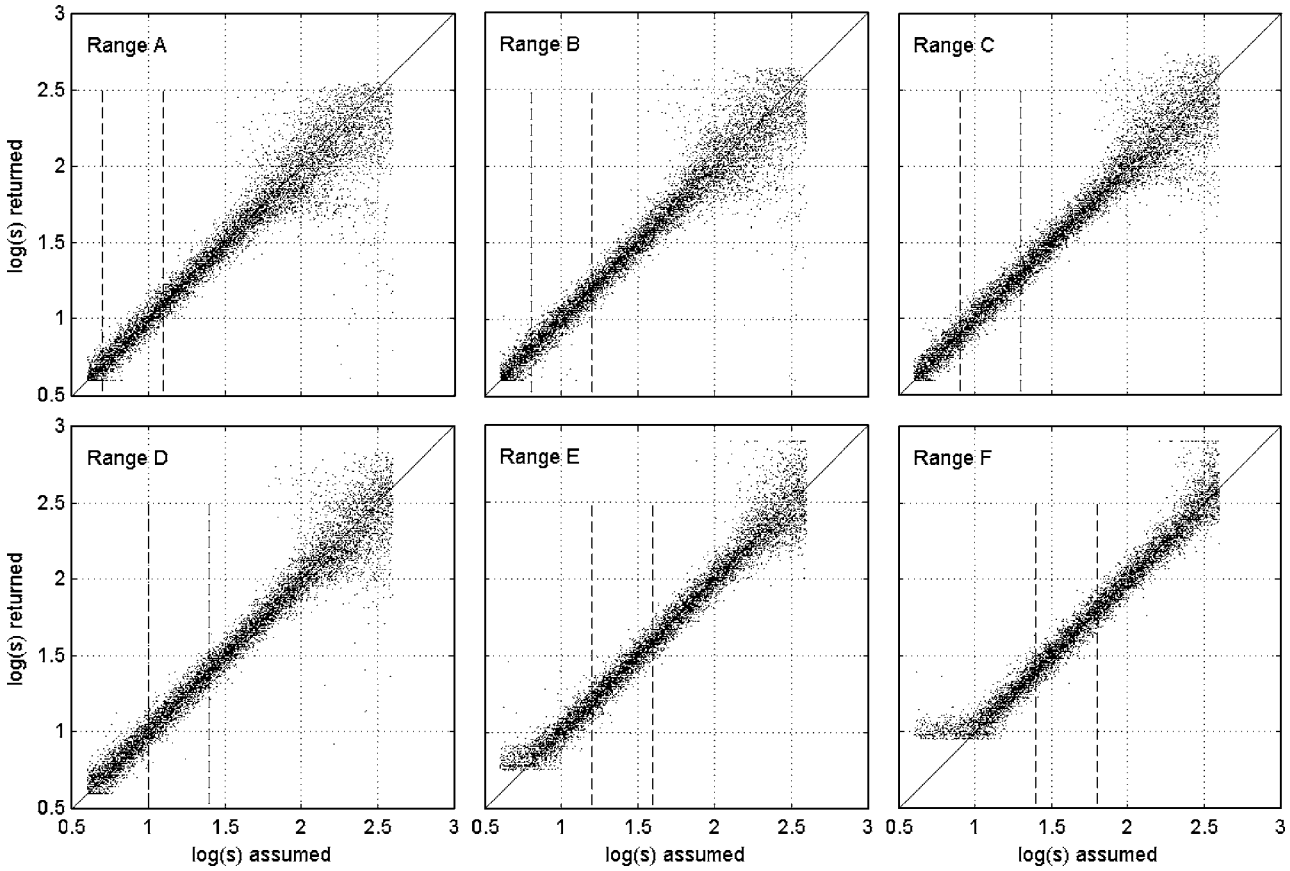

Fig. 8 Scatterplot with the assumed stray light value on the *x* axis, and the stray light value returned by the Monte-Carlo simulation on the *y* axis. The interval of stray light values simulated is uniformly distributed from $\log s = 0.6$ to 2.6. The figure shows the result of 8000 simulations. The vertical dashed lines indicate the (age dependent) 95% confidence limits of stray light values found in an average population.

and ESD shown in this figure were smoothed by a moving average with window size $n = 100$. For the first three simulations, shown in Figs. 7(a)–7(c), SD *does* follow ESD. The effect of sample size (300, 60, and 30, respectively) is very clear in these three figures. The fourth simulation, with insufficient range of trial levels, is shown in Fig. 7(d). For the larger ESD values in this figure, ESD deviates strongly from the $y = x$ line. ESD tends to overestimate the true SD for values larger than 0.1 (see also Fig. 5).

The second set of Monte-Carlo simulations tests the two-phase sampling scheme developed for stray light measurement in clinical practice, as explained in Sec. 2. This sampling scheme consists of an initial estimate of the ML estimation stray light level, with a relatively coarse sampling distance of 0.1 logarithmic units. The measurement is refined in a final phase with a sampling distance of 0.05 logarithmic units. The range used in the initial phase can be chosen by the operator, as summarized in Table 1. Measurement ranges A to F were used in the second set of Monte-Carlo simulations. As before, each simulation contains 8000 assumed stray light values, uniformly distributed from $\log(s) = 0.6$ to $\log(s) = 2.6$.

Figure 8 shows the returned stray light value as a function of the assumed values in the simulation. The two vertical dashed lines indicate the stray light intervals for which the ranges were intended. These correspond to the 95% confidence intervals of stray light values in the respective age group, as given in Table 1. The spreading of the results around the $y = x$ line shows that these intervals are rather conservative. Up to some distance (–0.2 and +0.5 logarithmic units respectively) outside these intervals, reliable measurements are obtained.

Figure 9 shows ESD as function of assumed stray light value. This figure shows quantitatively what interval of stray light values can be measured to a certain degree of accuracy in each range setting. For example, in range E it can be seen that this interval is $1.1 < \log(s) < 1.9$ for an accuracy of 0.07 logarithmic units. ESD values outside the usable interval rapidly increase to large (>0.1) values. The reason for this rapid increase is illustrated in the example given in Fig. 5. Mismatch between measurement range and stray light value causes the lowest confidence level used for ESD calculation not to be bound by the likelihood ratio function.

3.2 *Mismatch of PF*

The last series of Monte-Carlo simulations was used to approach the true field situation as closely as possible, with the GLARE study as a reference. An important aspect of these simulations is that the simulated PF differs from the fixed shape that is used in the ML analysis. The 12 PF that were used in these simulations are shown in Fig. 3. The stray light values assumed and the range settings in the initial phase were

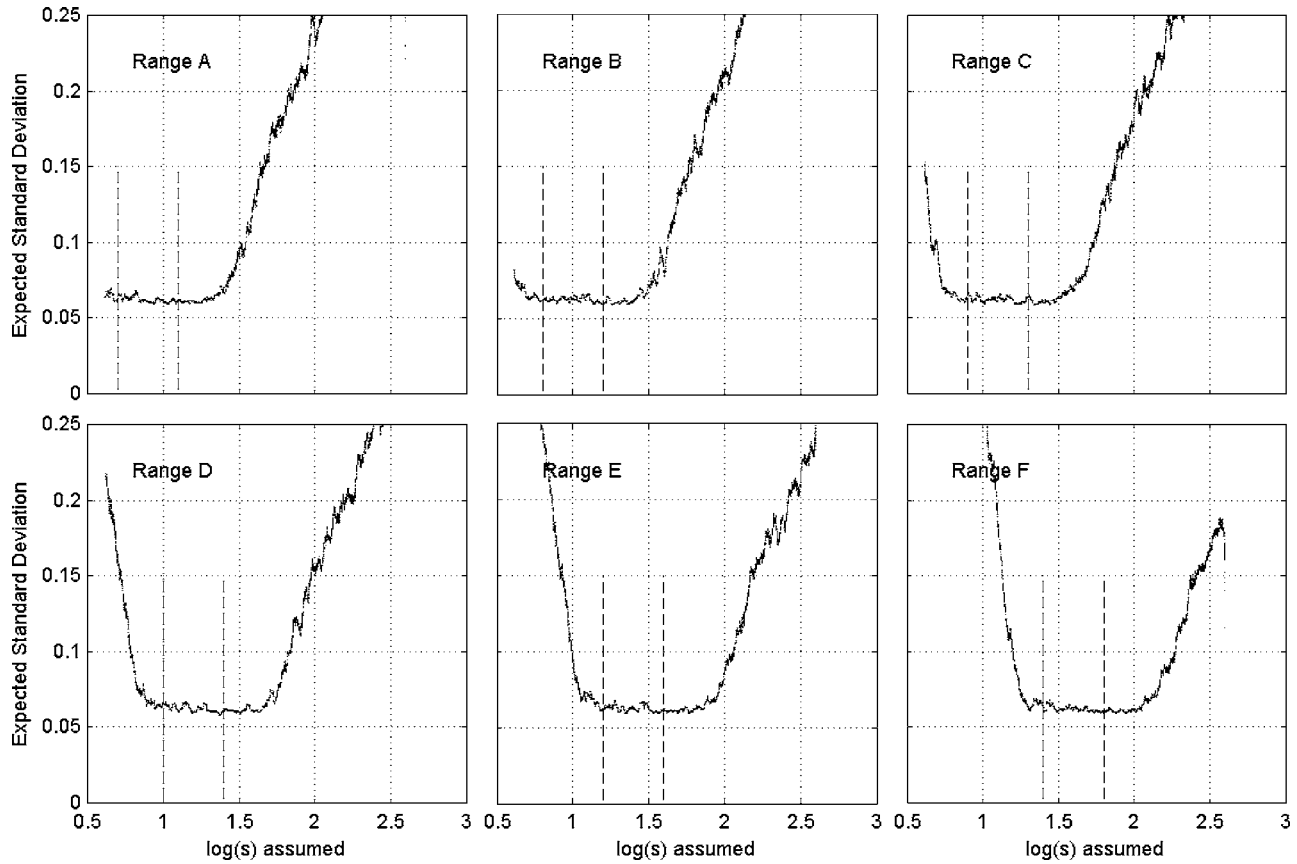

Fig. 9 ESD as function of simulated stray light value. The interval of stray light values simulated is uniformly distributed from log(s)=0.6 to 2.6. The figure shows the result of 8000 simulations. ESD values have been smoothed by a moving average (n=100). This figure can be compared with Fig. 8 that shows the raw simulated stray light values directly. Again, the vertical dashed lines indicate the (age dependent) 95% confidence limits of stray light values found in an average population.

taken from the GLARE study. So, each simulation contains the same 8230 stray light values and range settings. Figure 10 shows the returned stray light value as function of the assumed value. From left to right and top to bottom the PF is less steep.

The plots contain less than 8230 data points, because a limit value for ESD of 0.08 was used. Note that in the field ESD is used to accept or reject a measurement and redo a measurement if necessary. For a steep PF, almost all simulations resulted in acceptable ESD. This is the case for most of the 12 simulations in Fig. 10. However, for the most shallow PF (lower right) less than half of the simulated values had an ESD$<$0.08.

On average there is no significant difference between simulated and returned values. The largest systematic difference (only 0.020 logarithmic units) is found for the shallowest PF. On average, for a very steep PF (top left), ESD=0.055 and SD=0.034. So, for a very steep PF, ESD *over*estimates SD. For a very shallow PF (bottom right), ESD=0.068 and SD=0.107. So, for a very shallow PF, ESD *under*estimates SD.

Figure 11 shows the relationship between SD and ESD for the same simulations shown in Fig. 10. For very good observers (top row), almost all simulated measurements had an ESD smaller than the limit value that was set to 0.08. For average observers (middle row), the range of returned ESD values starts to increase, and the data tend toward the $y=x$ line. Note that for these observers, actual PF and assumed PF are virtually identical (see Fig. 3). For bad observers (lower row, right two plots), a smaller percentage of the measurements reached the limit value for ESD. The data lie above the $y=x$ line, indicating that ESD *under*estimated SD.

4 Discussion

None of the simulations have shown significant systematic differences between assumed and returned stray light values. However, this statement only holds when results with ESD higher than 0.08 are excluded. Since in practice this should be the case, no significant bias is expected on the basis of the MC analysis presented here. Using the 0.08 limit value for ESD, the largest systematic difference between returned and assumed stray light value is 0.02 logarithmic units, as obtained from the worst group of observers in Fig. 10 (lower right plot). With a random error (SD) of a CC test of about 0.05 logarithmic units, a systematic error of 0.02 logarithmic units can be considered acceptable.

4.1 *SD and Its Relation to ESD*

SD may be expected to be proportional to the reciprocal of the square root of the sample number, as shown in Figs. 6(a)–6(c). However, also sampling density may be expected

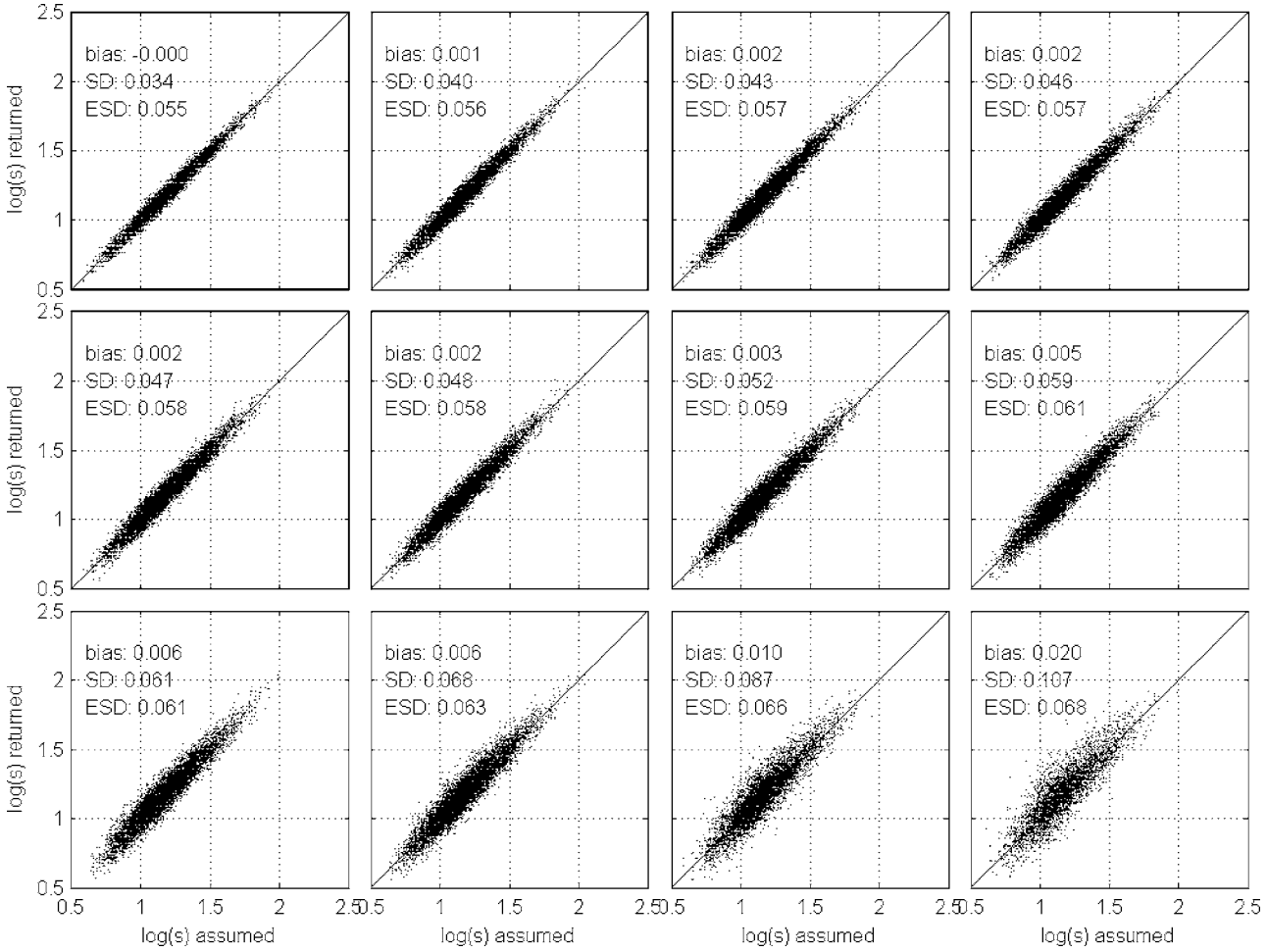

Fig. 10 Returned stray light value as function of the assumed value. The PFs used in the simulations equal those shown in Fig. 3, and differ from the PF used in the ML estimation.

to play a role. If sampling density is too coarse compared to the steepness of the PF, accuracy suffers. Other simulations, not presented in this paper, have shown that, for example, identical results are obtained with a spacing of 0.01 logarithmic units, and a five times repetition at 0.05 logarithmic units spacing. For a coarse sample spacing at 0.10 logarithmic units and with a 10 times repetition this no longer holds; some discretization is observed, faintly visible in Fig. 6(c). So, a safer choice in practice would be a spacing of 0.05 logarithmic units.

Figures 7(a)–7(c) show that ESD and SD correspond very well. So, the theoretical need for a large number of samples to use the likelihood ratio function as predictor for data reliability is easily met in practice. Even for the rather coarse sampling with a spacing of 0.10 logarithmic units, as shown in Fig. 6(c) and Fig. 7(c), the asymptotic conditions required in theory seem to be reached in practice.

For an inadequate range of test levels, as shown in Fig. 6(d), stray light values outside the sampling interval cannot be measured accurately. This is quite obvious, since most information about the stray light value of a subject is obtained from the transition from 0 to 1 responses near twice the stray light value. Trials presented at compensation levels (far) away from this transition carry little information.

ESD was found to represent the true SD in Figs. 6(a)–6(c) and Figs. 7(a)–7(c). However, as explained earlier, these simulations used impractical sampling schemes. When using the more practical two-phase sampling strategy, ESD represents true SD less precise, but it turned out to be a conservative (i.e., *safe*) estimate of SD: In this case, assumed stray light values lie within the test range chosen, ESD does represent SD. In this case, the stray light values are outside this interval, ESD rapidly increases and more so than SD.

So, in the case of inadequate sampling, ESD is a conservative value that overestimates the true SD to be expected. However, ESD calculation is based on an assumed shape of PF. Data from the GLARE study suggest that this assumption is invalid. This raised the question of what happens to the ESD value when there is a mismatch between the PF of an observer and the PF used in the ML analysis.

The upper row of Fig. 10 shows results from simulated observers with a PF that is steeper than the PF used in the ML analysis. In this case, the SD is lower than ESD. The lower row of Fig. 10 shows results from simulated observers with a PF ranging from similar to much shallower than the PF in the ML analysis. The corresponding SDs increase with shallower PF. For the two lower right cases, with very shallow PF, ESD clearly underestimates the true SD, which constitutes a cau-

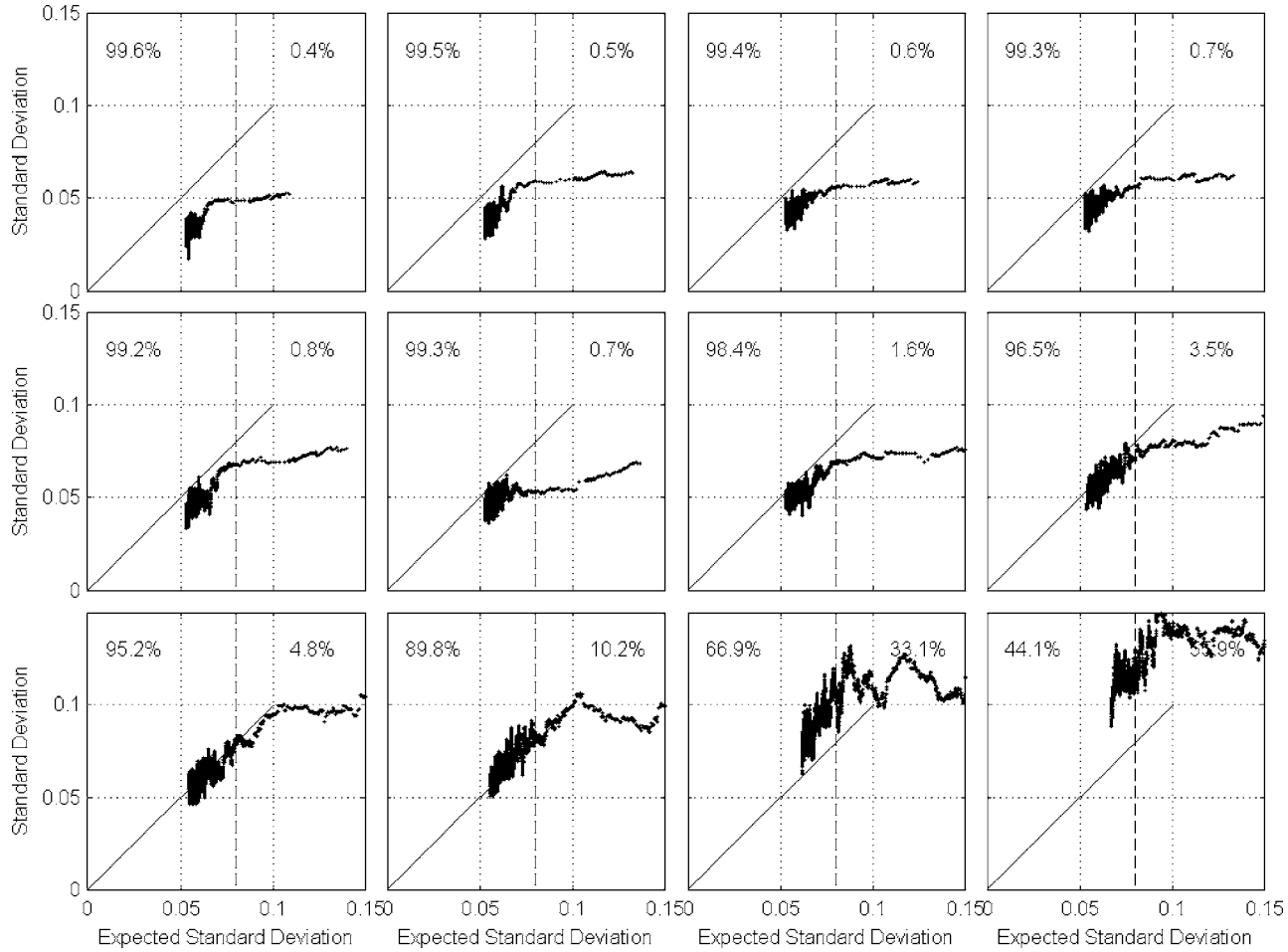

Fig. 11 SD as function of ESD for the same series of simulations shown in Fig. 10. The data have been smoothed by a moving average, with a $n=100$ window size. In the ideal case, ESD would equal SD, and all data would lie on the $y=x$ line. The dashed vertical line indicates the limit value of 0.08 for ESD that was used. The percentages give the distribution of data points above and below the ESD limit value.

tionary note. However, the difference is not large. The percentage of acceptable measurements ($ESD<0.08$) is quite low though in these cases (see Fig. 11).

4.2 *Effectiveness of the Two-Phase Sampling Scheme*

The CC measurement as implemented in practice consists of an initial phase of 12 samples spaced by 0.10 logarithmic units, and a final phase of 13 samples spaced by 0.05 logarithmic units, as illustrated in Fig. 2. The final phase has the samples placed around compensation levels that carry most of the information about the true stray light value of a subject. These samples are placed near the transition from 0 to 1 responses. The range of test levels in the initial phase can be chosen to be most efficient for the expected stray light value for the subject. Table 1 gives for each range setting an interval of stray light values. The effectively usable interval proved to be wider than the interval for which the range was intended. Figure 9 shows that the usable interval is about 1 logarithmic unit, slightly decentered toward the higher stray light values. This figure also indicates that within the usable interval, an SD of 0.06 logarithmic units can be expected, which is clearly acceptable in clinical practice.

To summarize, Monte-Carlo analysis was used to investigate properties of the CC method for the assessment of retinal stray light. In practical application, no significant bias is to be expected. The ESD value obtained in a CC test approximates true SD in the majority of cases. ESD is a conservative estimate if the sampling range is not chosen properly. Only for subjects with a very poor PF, showing $ESD>0.1$, ESD tends to underestimate true SD. In the vast majority of cases, the sampling strategy proved to be adequate, giving a SD between 0.1 and 0.03.

Appendix: Psychometric Function

The light the fovea receives in a CC test consists of two parts: light originating from the flickering annulus by the process of scattering, and light originating from the half fields the subject is looking at. Both lights correspond to certain luminances in the outside world (in the two half fields). The light originating from scatter (i.e., the stray light) corresponds to an outside luminance called equivalent luminance, L_{eq}.[4] For the stray light source used here (an annulus with a 1:2 ratio of inner and outer radius, see Fig. 1), $L_{eq}=0.0013sL_{src}$, with L_{src} being the luminance of the annulus (in its on phase) and s the

stray light value of an eye. For more details on equivalent luminance, stray light source illuminance and its relation to source geometry see Ref. 12. Now let us express the externally presented luminances L in the test fields as a fraction of L_{src} or to be precise as $S=L/(0.0013L_{src})$. The unit of S is $[\mathrm{deg}^2/\mathrm{sr}]$, which is (apart from a constant) dimensionless. By this choice of "*s* units," the luminances used can be compared directly to the s value of the subject, independent of L_{src}.

The two test fields are referred to as field a and b. Field a is never given compensation flicker; field b is given various amounts of (external) compensation luminance, S_{comp}, during the off phase of the stray light source in a test and none during the on phase of the source. The average luminance of fields a and b is kept equal by adding $0.5S_{comp}$ to field a in both the on and the off phases (of the stray light source). The luminances used are $Sa^{on}=s+0.5S_{comp}$, $Sa^{off}=0.5S_{comp}$, $Sb^{on}=s$ and $Sb^{off}=S_{comp}$, where S_a and S_b represent stimulation of the *retina*, corresponding to the sum of the (external) luminance of the test fields and the *equivalent* luminance of the light scattered from the straylight source. The perceived flicker strength in each of the test fields is given by their respective modulation depths: $MDa(S_{comp},s)=|(Sa^{off}-Sa^{on})/(Sa^{off}+Sa^{on})|$ and $MDb(S_{comp},s)=|(Sb^{off}-Sb^{on})/(Sb^{off}+Sb^{on})|$. The relative difference of these modulation depths, called modulation depth contrast (MDC), is consequently calculated as $MDC(S_{comp},s)=(MDb-MDa)/(MDb+MDa)$. Note that on a linear S_{comp} scale, $MDC(S_{comp},s)$ shows symmetry around s (see Fig. 1). On the logarithmic S_{comp} scale normally used in PF plots (e.g., see Fig. 2) this symmetry is not immediately noticeable.

A logistic function[13] was used as the basis for the PF of a CC task

$$P(S_{comp},s)=\lambda+(1-2\lambda)\left(\frac{1}{1+\exp\left[\dfrac{MDC}{MDC_c}\right]}\right),$$

where MDC_c is a critical value for MDC, and λ the lapsing rate describing nonperfect performance. An *a priori* choice for the PF was made with $MDC_c=0.156$ and $\lambda=0.05$. This PF was used for initial analysis of each individual measurement. The grouped population data were fitted with MDC_c free (fit results 0.070 to 0.52) and $\lambda=0.01$ (fixed).

References

1. L. Franssen, J. E. Coppens, and T. J. Van den Berg, "Compensation comparison method for assessment of retinal straylight," *Invest. Ophthalmol. Visual Sci.* **47**, 768–776 (2006).
2. T. J. T. P. van den Berg, "Importance of pathological intraocular light scatter for visual disability," *Doc. Ophthalmol.* **61**, 327–333 (1986).
3. M. A. Mainster and G. T. Timberlake, "Why HID headlights bother older drivers," *Br. J. Ophthamol.* **87**, 113–117 (2003).
4. J. J. Vos, "Disability glare—A state of the art report," *Commission Int. l'Eclairage J.* **3/2**, 39–53 (1984).
5. J. J. Vos and T. J. T. P. van den Berg, "Report on disability glare," *CIE Collection* **135**, 1–9 (1999).
6. D. B. Elliott and M. A. Bullimore, "Assessing the reliability, discriminative ability, and validity of disability glare tests," *Invest. Ophthalmol. Visual Sci.* **34**, 108–119 (1993).
7. W. R. Meacock, D. J. Spalton, J. Boyce, and J. Marshall, "The effect of posterior capsule opacification on visual function," *Invest. Ophthalmol. Visual Sci.* **44**, 4665–4669 (2003).
8. S. C. Schallhorn, C. L. Blanton, S. E. Kaupp, J. Sutphin, M. Gordon, H. Goforth Jr., and F. K. Butler Jr., "Preliminary results of photorefractive keratectomy in active-duty United States Navy personnel," *Ophthalmology* **103**, 5–22 (1996).
9. J. E. Coppens, L. Franssen, L. J. van Rijn, T. J. T. P. van den Berg, "Reliability of the 'compensation comparison' straylight measurement method," *J. Biomed. Opt.* **11**, 034027 (2006).
10. W. Q. Meeker and L. A. Escobar, "Teaching about approximate confidence regions based on maximum likelihood estimation," *Am. Stat.* **49**, 48–53 (1995).
11. B. Treutwein, "Adaptive psychophysical procedures," *Vision Res.* **35**, 2503–2522 (1995).
12. T. J. T. P. van den Berg, "Analysis of intraocular straylight, especially in relation to age," *Optom. Vision Sci.* **72**, 52–59 (1995).
13. H. Strasburger, "Converting between measures of slope of the psychometric function," *Percept. Psychophys.* **63**, 1348–1355 (2001).