Cross-validation and robustness of daylight glare metrics

Wienold, J.; Iwata, T.; Sarey Khanie, Mandana; Erell, E.; Kaftan, E.; Rodriguez, R. G.; Yamin Garreton, J. A.; Tzempelikos, T.; Konstantzos, I.; Christoffersen, J.

Total number of authors:
13

Published in:
Lighting Research and Technology

Link to article, DOI:
10.1177/1477153519826003

Publication date:
2019

Document Version
Peer reviewed version

Research

Corresponding author**:**
Wienold J., Interdisciplinary Laboratory of Performance Integrated Design (LIPID), EPFL, LE 1 111 (Bâtiment LE) Station 18 CH-1015, Lausanne, Switzerland,
Email: jan.wienold@epfl.ch

Cross-validation and robustness of daylight glare metrics

Wienold J. [a], Iwata T. [b], Sarey Khanie M. [c], Erell E. [d], Kaftan E. [d], Rodriguez R. G. [e], Garreton J. Y. [e], Tzempelikos T. [f], Konstantzos I. [g], Christoffersen J. [h], Kuhn T. E. [i], Andersen M. [a]
[a] LIPID, School of Architecture Civil and Environmental Engineering, EPFL, Lausanne, CH
[b] , Department of Architecture and Building Engineering, Tokai University, Hiratsuka, JP
[c] , Department of Civil Engineering, Technical University of Denmark DTU, DK
[d] , Ben-Gurion University of the Negev, Israel
[e] , INAHE Institute of environment, habitat and energy, CONICET, Mendoza, AR.
[f] , School of Civil Engineering and Center for High Performance Buildings, Purdue University, West Lafayette, Indiana USA
[g], Charles W. Durham School of Architectural Engineering and Construction, University of Nebraska–Lincoln, Omaha, Nebraska, USA
[h] VELUX A/S, Hoersholm,DK
[i] Fraunhofer Institute for Solar Energy Systems ISE, Freiburg, D

Abstract

This study evaluates the performance and robustness of twenty-two established and newly proposed glare prediction metrics. Experimental datasets of daylight-dominated workplaces in office-like test rooms were collected from studies by seven research groups in six different locations (Argentina, Germany, Denmark, Israel, Japan and USA). The variability in experimental setups, location and research teams allowed reliable evaluation of the performance and robustness of glare metrics for daylight-dominated workplaces.
Independent statistical methods were applied to individual datasets and also to one combined dataset to evaluate performance and robustness of the twenty-two glare metrics. As performance and robustness are not established in literature, we defined performance as: 1) the ability of the metric value to describe the glare scale (evaluated by Spearman's correlation), and 2) the ability of the metric to distinguish between disturbing and non-disturbing situations (evaluated by diagnostic ROC-curve-analysis-tests). Furthermore, we defined robustness as the ability of a metric to deliver meaningful results when applied to different datasets and to fail as few as possible statistical tests. Average Spearman correlations in the range of 0.55-0.60 as well as average prediction rates to distinguish between disturbing and non-disturbing glare of 70-75% for several of the metrics show that the results of the metrics are in general trustable - therefore, the poor performance reported in some studies cannot be confirmed. Also, the cross-validation results show that metrics considering the saturation-effect as a main effect in their equation perform better and more robustly in daylight-dominated workplaces than purely contrast-based metrics or purely empirical-derived metrics. The

results indicate that the Daylight Glare Probability (DGP) delivers the highest performance amongst the tested metrics and was also found to be the most robust one.
Future research should aim to optimize the terms of glare equations which combine contrast and saturation effects, for example DGP, PGSV or UGR_{exp}, to achieve metrics performing also reliably in dimmer lighting conditions than the ones explored in this study.

Keywords
Discomfort glare, Daylight, HDR imaging techniques, glare perception, user assessments, glare protection

Introduction and objectives

The avoidance and prevention of glare is an important issue for the design and operation of buildings to guarantee a comfortable visual environment for occupants. The existence of reliable glare metrics is therefore a necessity. In the past decades façade installations have advanced in order to balance different and contradicting criteria[1] such as, amongst others, view out, daylight provision, solar and glare protection, ventilation, sound protection and aesthetics. Many of the resulting façade designs lead to rather complex light distributions, making it challenging for glare metrics to produce reliable results. Most of the existing metrics have been developed under specific, non-general boundary conditions with limited variations. As a result, several studies [2–4] reported poor overall performance of existing glare metrics. Some of these studies therefore proposed new or modified daylight glare or visual discomfort metrics, based on the acquired user assessment data. A common restriction of these studies is the limited variation of the luminous environment while developing metrics or modifying existing ones, as well as the limited amount of data points, something unavoidable in such experiments. Often window sizes, seating position and viewing directions are not changed during experiments. While a metric might perform better under a certain condition, it may fail for conditions which are significantly different.
Hence, the question remains: how well do glare metrics perform when the lighting conditions are different from the ones under which they were developed?
The objective of this study is to evaluate the performance and robustness of established and newly proposed glare prediction metrics by using data-sets from six studies that were not used to develop the metrics themselves. An additional dataset from Germany and Denmark, which was used to develop the DGP metric, is also used for training the metrics (see section 1.1). The dataset used as a whole for this cross-validation study was acquired by different research groups in Germany, Switzerland (though experiments were conducted in Germany), Denmark, Japan, Israel, Argentina and USA. This variability in locations, climatic zones and research teams allows the evaluation of the performance and robustness of glare metrics much more reliably than using only a single data-set from one research team. To limit confounding factors to a minimum, only studies which conducted the experiments under daylight in controlled environments (office-like test rooms) were eligible for the present cross-validation effort.

1. Selection of glare metrics

Having a series of large and diverse datasets available to investigate the performance and robustness of glare metrics, it is appealing to evaluate as many potential metrics as possible. However, statistical reasons limit the number of metrics that can be investigated. One of the main reasons is the risk of a type I error ("false alarm") which increases linearly with the number of analyses, relative to the number of tested independent metrics. When several, independent metrics are investigated it requires an adjustment of the significance level. E.g. for 2000 metrics, the significance level needs to be adjusted from 0.05 to 0.000025 before accepting the results (see section 2.5.2, which discusses the Bonferroni correction of the significance-levels). Ignoring this rule can lead to random results (type I error). On the other hand, if this adjustment is applied and a large number of metrics are tested, such a low p-value is hard to achieve for any metric – all of them would fail the significance threshold. For this reason, we restricted the number of metrics evaluated to the most relevant ones, which counts down to 22 metrics. In this selection, we have long-time established metrics like the Daylight Glare Index DGI, which was the first glare metric dedicated to evaluating daylit environments. We also selected several well-established glare metrics in the field of electric lighting, which are mentioned in standards and CIE-documents (UGR, CGI). Furthermore, we considered metrics that were recently published or mentioned in publications or that are revisions of long-term established metrics. DGP_{mod}[5], a recent modification of DGP to be used in cases of direct sun visible through shading fabrics, was discarded in the evaluation, as this metric was extracted from a dataset, which included the US-Fabrics dataset used in this study. Including DGP_{mod} would thus violate the restriction of not using datasets used to develop the evaluated metrics (more in Section 2). Since the US-Fabrics dataset in itself offered unique circumstances (sun through shading fabrics), which were crucial to the completeness of examined conditions in this study, the authors decided to include the dataset, but not the DGP_{mod} metric for the statistical evaluations. However, in the discussion section 4.3 we compare the core performance values of DGP_{mod} with the unmodified DGP metric to evaluate differences between the metrics.

The complete list of evaluated glare metrics and their related main publication can be found Table 1, their equations and/or definitions are listed in the supplementary material.

Table 1: Overview of the investigated metrics (22 in number) including references, where the metrics were either developed, suggested or discussed. The equations and/or definitions of the metrics are listed in the supplementary material.

#	Name	Variable Name	ref.	#	Name	Variable Name	ref.
1	CIE Glare Index	CGI	6,7	**12**	Median Luminance of Lower Window (<2m height)	$L_{med\_lowerwin}$	3
2	Daylight Glare Index	DGI	8,9	**13**	Median Luminance of Window	L_{med_win}	3
3	Modified Daylight Glare Index	DGI_{mod}	10	**14**	Position Index Weighted Average Luminance of Image	L_{pos_avg}	
4	Daylight Glare Probability	DGP	11,12	**15**	Standard Deviation of the Luminance of the Window	L_{std_win}	3
5	Direct Illuminance	E_{dir}	5	**16**	Perceived Glare Level for typing task	PGL	13
6	Illuminance at Eye Level	E_v	11,14	**17**	Predicted Glare Sensation Vote	PGSV	15,16
7	Glare Sensation Vote	GSV	17	**18**	Predicted Glare Sensation Vote (saturation glare)	$PGSV_{sat}$	15
8	Average Luminance in 40° Band	L_{40band_avg}	3	**19**	Unified Glare Probability	UGP	4
9	Average luminance in Image	L_{avg}	2	**20**	Unified Glare Rating	UGR	7
10	Average Luminance of Window	L_{avg_win}	3,11	**21**	Experimental Unified Glare Rating	UGR_{exp}	10
11	Median Luminance of Image	L_{med}	3	**22**	Visual Comfort Probability	VCP	18

2. Methodology

The data from seven different previously published studies were collected, screened, cleaned and evaluated following a common procedure. The studies were conducted in six different countries in different continents and climatic zones. All of them were conducted in controlled office-like environments where daylight was the main light source. In all studies, user assessments were accompanied by the acquisition of HDR images, which were taken either at the position of the subject's head or very close to it. All HDR images were screened for validity (e.g. pixel overflow) for this study. In few cases where pixel overflow occurred, the images were either corrected or deleted from the data (see 2.2). Five of the datasets used the same 4-point rating scale. The rating scales of the other two datasets were transformed to the 4-point scale of the other five datasets (see 2.4).

Testing performance and robustness of glare metrics requires various steps of data preparation and application of statistical tests. As a general rule and to avoid any biasing of the results by model training or metric-development-data, the statistical tests within this study were only applied to datasets that were neither used for the development of the metrics nor for the derivation of any borderline value (a borderline is the value separating two categories of a subjective ordinal rating scale, e.g. the borderline between noticeable and disturbing glare).

For the evaluation of the performance of the metrics, Spearman correlation and AUC were used. The robustness was tested by the squared distance and the True Positive Rate (TPR) and True Negative Rate (TNR). The overall methodology is illustrated in Figure 6 and described in detail in the following paragraphs 2.1 - 2.5.

2.1. Overview of test locations

Experimental data from seven different studies are used for this research. The studies were selected to have a high variability of tested shading systems and a high variability on climate/cultural background. In addition, the studies fulfilled following quality criteria: High control over experimental conditions, high reliability of the underlying data (e.g. no systematic error in images), and avoidance of reflections on computer screens.

All experiments were conducted under daylight in controlled office-like test rooms. A summary of the test locations is given in Table 2. For more detailed descriptions, please refer to the supplementary materials or to the original published studies.

Table 2: Overview of the studies, their experimental setup, equipment and examined subjects. The no. of cases specifies the total number of situations evaluated per study. The maximum amount evaluated by one test person was restricted to six per day for this study and varied between one (DE-Gaze), three (AR-DEO, DE-DK-Ecco, JP-Office and IL-DayViCE), four (DE-Quanta) and six (US-fabric). Data which was used to develop DGP is not used for any evaluation of the metrics. It is only used for the training of the metrics (see 1.1).

Study	Research team	ref.	place	Window size	Shading/ Daylight system	Test-persons	Cases	Cases Non-development
AR-DEO	Yamin Garretón Rodriguez	19	Mendoza, Argentina	S	Glazing without shading	27	55	55
DE-DK-Ecco	Wienold Christoffersen	11	Freiburg, Germany Copenhagen, Denmark	S M L	White Venetian blinds Specular Venetian blinds Transparent foil system	59 24	229 130	14 0
DE-Gaze	Sarey Khanie	20	Freiburg, Germany	L	Glazing without shading	95	95	95
DE-Quanta	Wienold	21	Freiburg, Germany	L	White Venetian blinds 2 types fabric roller shades	49	196	196
IL-DayViCE	Erell Kaftan	22	Sde Boqer, Israel	L	Light shelf Transparent foil system Glazing without shading Venetian blinds	59	151	151
JP-Office	Iwata	15	Tokyo, Japan	L	White Venetian blinds	72	162	162
US-Fabric	Tzempelikos Konstantzos	5	West Lafayette, USA	L	14 types fabric roller shades	35	141	141
Total data						**420**	**1159**	
Total non-development data						**337**		**814**

With:
Window size S: Small (glazing fraction: <40% of facade)
Window size M: Medium (glazing fraction: >=40% and <70% of facade)
Window size L: Large (glazing fraction: >=70% of facade)

2.2. Quality and consistency of experimental data

To ensure reliable basis-data for the analysis, the entire HDR image-dataset was screened and checked. One major and possibly biasing issue in HDR-imaging and glare evaluation is the risk of pixel overflow when trying to measure high luminances (e.g. sun, reflections of sun, sun seen through fabric shadings). Another potential issue is, in cases where the camera is not at the subject's eye position, that the camera "sees" a significantly different situation than the subject (e.g. the sun disk is hidden by a frame of the window, where the user can see the sun or vice versa). All images from these studies were manually checked and the respective cases removed from the dataset. The pixel overflow was checked by comparing the illuminance measured besides the lens with the illuminance derived from the HDR image. As quality criterion, deviations of more than 25% between the measured and derived illuminance were considered unacceptable. Details of the selection, correction and quality process can be found in the supplementary material. The studies DE-Gaze and JP-Office had no illuminance-sensor installed besides the camera, therefore another procedure was applied to these images which is described in the supplementary material. The overall quality of the HDR-cameras used can be seen in Figure 1. For the unchanged images we calculated bias, normalized bias, root mean squared error RMSE and normalised root mean squared error NRMSE (for equations see supplementary material). For all studies the normalised bias was less than 5% - respectively less than 10% for the NRMSE - indicating that the quality of the images is good enough to conduct a validation study.

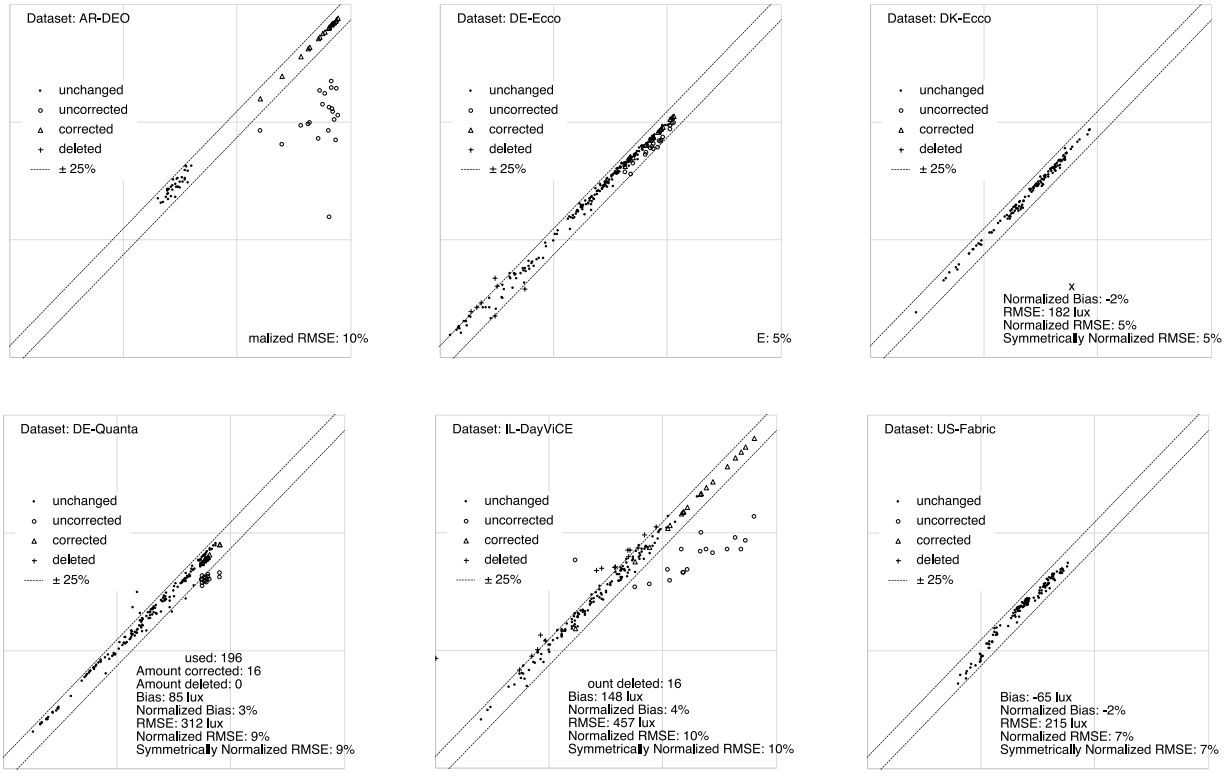


Figure 1: Comparison of measured and image-derived illuminance for the different datasets

2.3. Data overview

The individual studies were conducted with different façade settings ranging from no shading with sun in the field of view to small window shaded with venetian blinds, leading to a large variability of light distributions in the overall dataset from less than 100 lux to more than 80000 lux. This variability is illustrated in Figure 2, where the distributions of the illuminance levels at the eye and the average luminance of the window are shown in a violin plot for all the underlying studies. As the average illuminance values at eye level for all studies are above 1800 lux, it means all the experimental setups can be considered as daylight-dominated workplaces. Dimmer lighting conditions, as they occur e.g. in open-plan offices or position further away from the façade, are only marginally represented in the datasets and therefore the results may be extrapolated to these conditions only with caution.

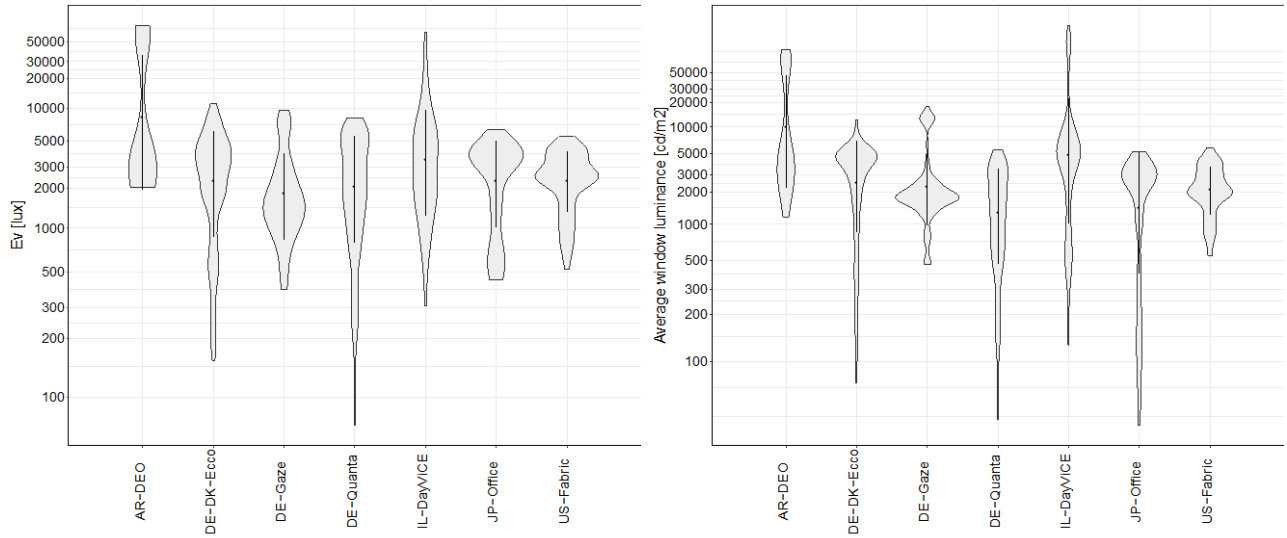


Figure 2: Violin plot showing the distribution of the illuminance at eye level E_v (left) and the average luminance of the window L_{win_avg} (right) for the studies. The dots in the figure are showing the median values, the vertical line the 50% percentile.

2.4. Subjective assessments – scales and mapping

The underlying glare-studies and their experiments were independently designed and, as a result, the subjective assessments were conducted with different procedures and scales. While the validity of the application of glare scales such as de Boer's scale [23,24] is under discussion, the current differences in scales might influence the glare prediction results [25]. More importantly, it makes a common evaluation such as the one being performed in this study difficult. In order to overcome this limitation, we needed to map the data to a single basis scale.

Out of the seven studies, five use the same four-point scale introduced by [26], which was found to be consistent [27] when compared to a linear glare scale (Cronbach alpha = 0.90). Moreover, the simple structure of the itemized four-point scale using "imperceptible", "noticeable", "disturbing" and

“intolerable” as a degree of glare descriptor reduces the confusion when the questionnaires are being initially filled out. We hence used this four-point Likert scale as basis-scale. The two studies, JP-office and IL-DayViCE, needed to be mapped to the basis scale in order to have a common data basis. The JP-office study used a linear scale with marks at the borderline between perceptible, acceptable, uncomfortable and intolerable glare. The IL-DayViCE study used a 5-point Likert scale from “not at all” to “very much” as responses. **Table 3** shows the glare- perception questions asked in each study and also illustrates how these two scales were mapped to the four-point scale.

Table 3: Glare questions, used subjective scales and mapping of the scales for the IL-DayViCE and JP-office study.

Study	Question	Scale & Mapping
AR-DEO	When reading the text on VDT, please mark the degree of glare you experienced from the window and the shading device.	☐ ☐ ☐ ☐ imperceptible noticeable disturbing intolerable
DE-DK-Ecco	When reading the text on VDT, please mark the degree of glare you experienced from the window and from shading device.	☐ ☐ ☐ ☐ imperceptible noticeable disturbing intolerable
DE-Gaze	Please specify the degree of glare you experience at the moment.	☐ ☐ ☐ ☐ imperceptible noticeable disturbing intolerable
DE-Quanta	When reading the text on VDT, please mark the degree of glare you experienced from the window and from shading device.	☐ ☐ ☐ ☐ imperceptible noticeable disturbing intolerable
IL-DayViCE	Were you bothered by glare (strong light) while copying text to the screen?	not at all … very much ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ imperceptible noticeable disturbing intolerable
JP-Office	When you look at the window, please mark the degree of glare you experienced from the window and the shading device.	just perceptible / just acceptable / just uncomfortable / just intolerable ☐ ☐ ☐ ☐ imperceptible noticeable disturbing intolerable
US-Fabric	Grade the visual discomfort (glare level), if any, that you experienced overall (any type of visual discomfort, bright objects, high overall brightness, contrast, reflections, shades, etc.) during your stay in the artition, considering that this situation can happen for varying amounts of time in your regular office.	☐ ☐ ☐ ☐ imperceptible noticeable disturbing intolerable

An overview about the relative occurrence of the subjective ratings can be seen in **Figure 3**.

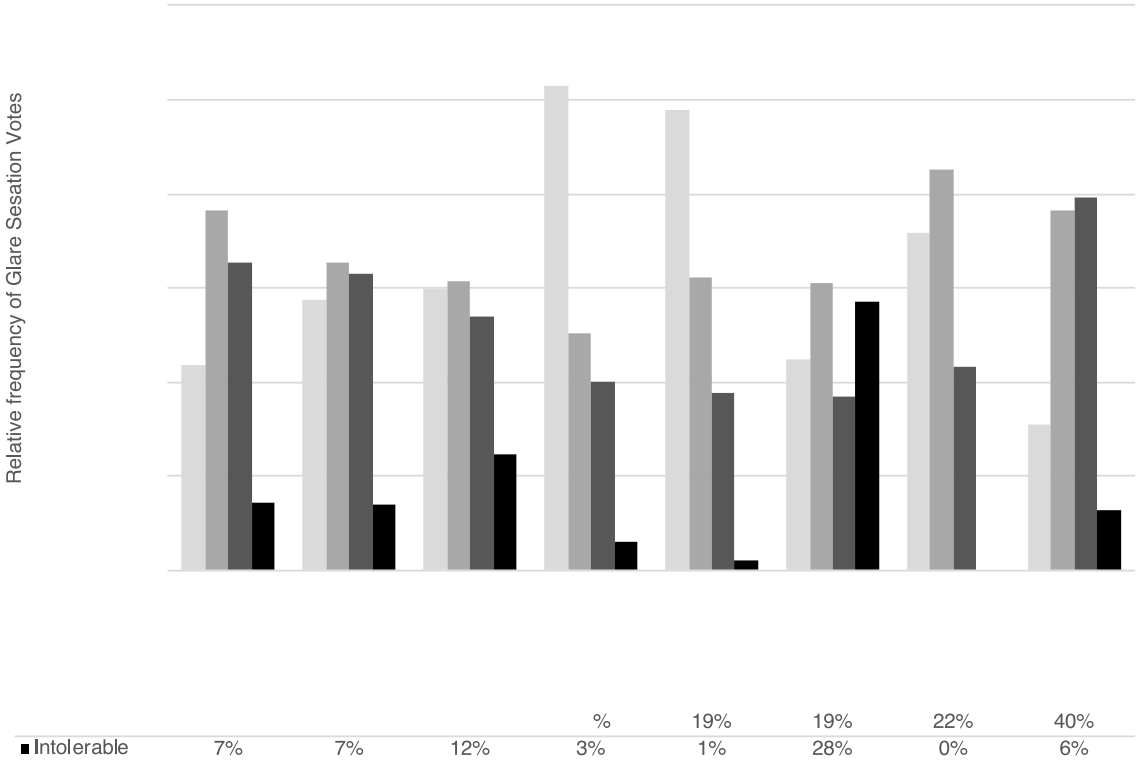


Figure 3: Relative distribution of the subjective responses for the underlying data

2.5. Statistical methods

The main goal of this study is to investigate performance and robustness of glare metrics. As of yet, there is no globally accepted statistical method available to evaluate the performance or robustness of glare metrics. But what does "performance" mean and how can we define robustness – both in relation to the evaluation of glare metrics?

To approach a reliable performance-evaluation, we can formulate two important questions:

1. How well does the metric describe the glare scale?
2. How well can the metric distinguish between disturbing and non-disturbing situations? Or in other words: Can a metric describe the probability that building occupants are disturbed by glare?

To answer the first question, the Pearson correlation has been used in some studies (e.g. [2,3]) for the evaluation of glare metrics. The Pearson correlation, however, only delivers reliable results for numerical or equidistant ordinal data, which is not typical of the subjective rating scales for glare. Here, since our data is of ordinal nature, we apply Spearman's correlation, which is accepted as an appropriate statistical approach for such data [32].

The second question can be answered by diagnostic tests, which are well known mainly in the field of medical research and have been introduced to the glare analysis by [33]. In our study, we applied the

Area Under the Curve (AUC, see 2.5.3.1) and the ROC Square Distance SqD (see 2.5.3.2) to evaluate the overall performance.

Robustness of the glare metrics and how to measure it in the context of glare metrics has not yet been defined in literature. In a general sense, we can define robustness as the ability of a metric to deliver meaningful results when applied to different datasets and to fail as few as possible statistical tests. To be more specific, this can be approached by answering the following questions:

a) Do glare metrics fail statistical tests when applied to multiple/different datasets? If yes, how often? Fewer failings indicate a more robust metric.
b) What is the minimum "detection-rate" delivered by the different metrics when applied on different datasets? A detection rate is defined here as the rate of correct predictions distinguishing between noticeable and disturbing glare for a universally derived disturbing-borderline-value.
c) Does the disturbing-borderline value vary when derived from the different studies, and by how much? The smaller the difference, the more robust a metric can be considered in identifying a disturbing situation caused by daylight glare. This evaluation also answers whether different stimuli among studies (which is the case here with the very different set-ups and locations) lead to similar borderline-values.

Details of the applied methods are described in 2.5.1-2.5.5.

2.5.1. Spearman rank correlation

The Spearman rank correlation ρ is a non-parametric test to measure the strength of the relationship between paired variables, in our case between the subjective ratings on the 4-point scale and the numerical metric values. The underlying independent variables don't need to be of numerical or equidistant-ordinal nature[32]. For the interpretation of the values, Cohen considers a ρ of > 0.5 as large effect size and > 0.3 as medium effect size [34]. Ferguson suggests more strict effect size-thresholds in [35]. However, comparing the effect size thresholds from Cohen and Ferguson applied on the Spearman correlation for the underlying data with the AUC-evaluation (see 2.5.3.1) and its interpretation of Hosmer-Lemeshow[36], the interpretation of Cohen is more consistent and is therefore used for this study. The significance levels for rejecting the 0-hypothesis is corrected according to Bonferroni (see 2.5.2).

2.5.2. Bonferroni correction of the significance-levels

A Bonferroni correction of significance values should be applied when multiple statistical tests of a hypothesis are performed to keep the risk of a type I error on the same level as if there would be only one test applied [37]. Multiple applications of tests increase the probability of a random result linearly, at least when the tested variables are independent from each other. To account for this, the Bonferroni correction is applied where significance levels (α) are divided by the amount of repeated tests (the amount of metrics). Thus, we consider a α-value of $\alpha = 0.05/22 = 0.00227$.

2.5.3. Diagnostic statistics

Diagnostic statistics are well established in medical research and are, in that field, often applied to evaluate if a diagnostic method is able to predict a certain medical disease. These statistical tests can also be applied to other disciplines and were introduced to glare evaluations in [33]. The main basis for these kinds of evaluations are datasets with binary dependent variables. For our data, the categorical

variable (=glare sensation vote) is converted into a binary variable, for example *non-disturbing (imperceptible and noticeable glare) ⇔ disturbing (or stronger glare).* These binary variables are then analysed using the receiver operating characteristic curve (ROC curve), where the true positive rate TPR (also called sensitivity) is plotted against the true negative rate TNR (also called specificity) for different critical values of the respective glare metric. TPR corresponds in our study to the prediction rate of disturbing glare and TNR corresponds to the prediction rate of no or non-disturbing glare.

In Figure 4, sample ROC-curves are shown to illustrate this diagnostic analysis. The curves illustrate the ability of the glare metrics to discriminate between disturbing and non-disturbing glare. The curves can be analysed in various ways and will be explained in 2.5.3.1-2.5.3.3.

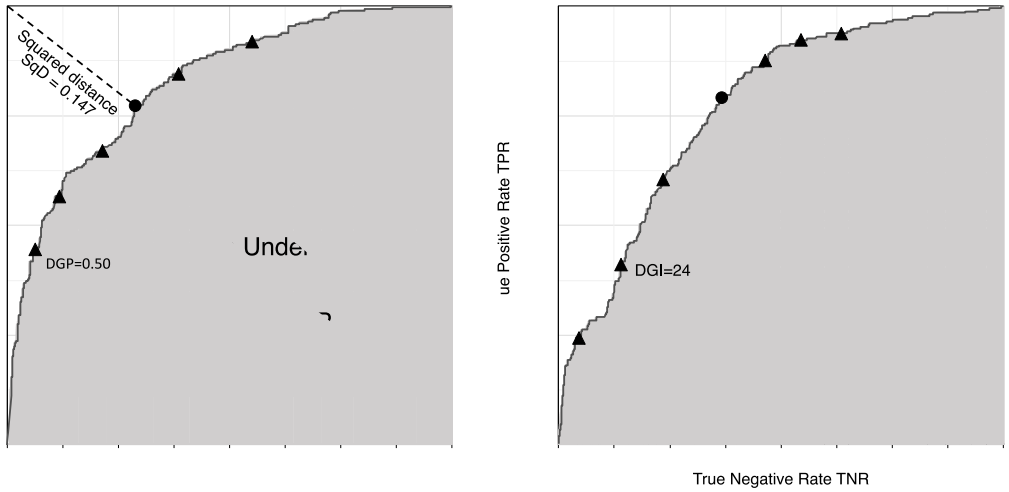


Figure 4: Example ROC curves for DGP and DGI for the combined dataset. Important critical values / borderline values of the metrics are displayed as triangles and dots. An ideal metric would touch the upper left corner (TPR=TNR=1) for ideally chosen critical values, which would mean a perfect prediction. The analysis of the squared distance SqD can be used to evaluate how well a metric can predict glare situations for a given borderline value. The SqD can also be used as basis for the determination of meaningful borderline values. The Area Under the Curve AUC describes the general ability of a metric to discriminate between disturbing and non-disturbing glare.

2.5.3.1. Area under the curve AUC

The area under the curve AUC describes the general ability of a metric to discriminate between disturbing and non-disturbing glare. It is a summary measure of the accuracy and is used in this study as a performance test. For the interpretation of the values, Hosmer-Lemeshow [36] describes an AUC ≥ 0.7 as an acceptable discrimination and an AUC ≥ 0.8 as an excellent discrimination for the binary data. Further, [38] interprets an AUC < 0.6 as fail and 0.7 > AUC ≥ 0.6 as poor.

2.5.3.2. Squared Distance SqD

The squared distance value SqD is used to analyse the ROC curve and is the squared distance from the curve to the upper-left corner, where the True Positive Rate TPR and True Negative Rate TNR are equal to 1 (see Figure 4). The smaller the SqD value, the better a metric is performing, with an ideal value at 0. In that case, the metric, together with the corresponding critical value / borderline

value, would reliably predict 100% of the disturbing glare situations and also 100% of the non-disturbing glare situations.
Typically, this distance is used to determine the borderline value, which is the best compromise between disturbing and non-disturbing glare. But this value can also be used as a performance indicator, when applied to a pre-determined borderline value (e.g. borderline value determined by another dataset or study). If this method is applied to different studies, the maximum value indicates the robustness of a metric.

2.5.3.3. True Positive Rate TPR and True Negative Rate TNR

While the True Positive Rate TPR and True Negative Rate TNR are very intuitive (TPR corresponds the prediction rate of disturbing glare situations and TNR corresponds to the prediction rate of no or non-disturbing glare situations), their combined evaluation is essential for meaningful interpretation of performance. A very high TPR could be reached by having a very low borderline value, causing a low TNR, which means the metric is over-predicting glare. The opposite happens when the borderline value is too high: TNR is very high then and glare under predicted. This characteristic behaviour can be used to evaluate the robustness of the metrics. Also, reporting an average TPR and TNR across studies could lead to wrong interpretations when the ratio between TPR and TNR is changing across studies. In that case a poor prediction rate in one study could be compensated by a high one in another. And this would happen in reverse order for the second prediction rate.
Therefore, we use these average values only as robustness indicators by evaluating the minimum prediction rates across the studies. The higher this rate is, the more robust is the metric. Also, we classify values below a minimum level of 50% (equals random level) as a failing of a test in our "failing analysis" in section 2.5.5.

2.5.3.4. Definition and determination of borderline values

Borderline values are metric values discriminating the (binary) dependent variable (user's perception). In non-lighting related research, typically the borderline-values calculated by diagnostic tests are also called "cut-off-values". An often-used borderline value in lighting research is the BCD, which is the borderline between comfort and discomfort. The BCD originated from the experiments by Luckiesh and Guth [39] and is mostly related to a semantic scale, which differs from the scale used for this study. For that reason, the borderline-values in this study are named differently. We used the following naming convention:

- BIN: Borderline between *imperceptible glare* ⇔ *noticeable glare*
- BND: Borderline between *noticeable glare* ⇔ *disturbing glare*
- BDI: Borderline between *disturbing glare* ⇔ *intolerable glare.*

In general, the relation between semantic scales and metric values is not consistent in literature [7,9,33,39–42]. Comparison of semantic scales used in the different metrics is beyond the scope of this study. Therefore, a direct comparison between "well accepted" BCD-values and the calculated borderline values of this cross-validation study should only be done with caution – especially for metrics mostly used for electric lighting (like CGI, UGR), which were developed using the de Boer's scale, which is different to the scale used for this cross-validation study.
There are several methods to determine an optimal borderline value, but none of these are perfect [43], so researchers typically select one of the methods for their study. For our study, we decided to apply the three most used approaches and use the average value of the three. The first method minimizes

the distance of the ROC-curve to the upper left corner (minimization of SqD). The second method maximizes the Youden index, which is defined as TPR+TNR-1. Graphically this corresponds to a maximum vertical distance between the diagonal and the ROC-curve. The third method is fitting a maximum sized square under the curve which results in the point TPR = TNR on the ROC-curve. The three methods are illustrated in Figure 5.

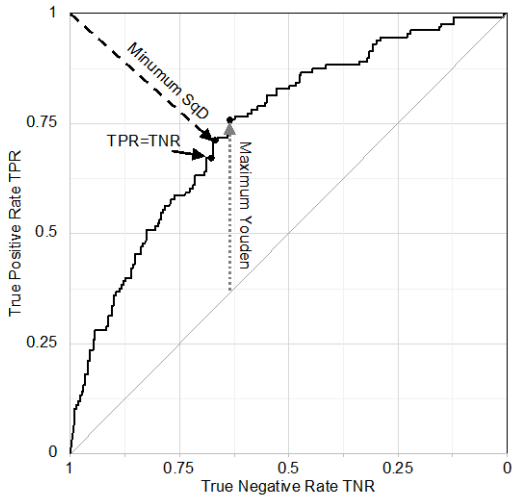


Figure 5: Three methods to determine the optimal borderline value: Minimizing the SqD, maximizing Youden index or Point on the curve, where TPR equals TNR. All methods are implemented in the evaluation and the average value of these 3 methods is used.

2.5.4. Logistic-regression

The logistic regression is a probabilistic prediction method, which is applied to dichotomous dependent variables. It is based on the hypothesis that the probability P is continuously increasing with a rising independent value following an S-shape. The regression fits the coefficients a and b of the equation (1) to the data. The p-value is used for the robustness analysis, comparing it to the Bonferroni adjusted significance level.

$$P = \frac{e^{a+bx}}{1+e^{a+bx}} \quad (1)$$

2.5.5. Robustness based on failure rate

The failing of significance tests is used in this study to evaluate the robustness of the already trained (see 2.7) metrics. The following tests are applied:

1. Spearman: The p-value of the Spearman-rank-correlation is compared with the Bonferroni-corrected significance value. Applied to each study separately.
2. Logistic regression: The p-value of the logistic-regression is compared with the Bonferroni-corrected significance value. Applied to each study separately.
3. TPR and TNR: The average TPR and TNR for each study is compared to a threshold of 0.5. A value of 0.5 or lower equals to a random result.
4. AUC: The AUC value is compared to a threshold of 0.6, which is interpreted as “fail” [38]. Applied to each study separately.

2.6. Dataset preparation

Following the principle of not using any development data, it was necessary to treat the data in two ways, depending on the statistical tests applied:

i. For statistical tests **not using borderline values** in the evaluation:
Here, all the data except development data is used. Each study results in one dataset, six in total. A seventh combined dataset is created containing all data from the six studies in order to have one dataset with a larger bandwidth of lighting situations.

ii. For statistical tests **using borderline values** in the evaluation:
The derivation of borderline values is treated as "training" of the glare models. The application of statistical tests is defined as "testing". For these two different phases ("training" and "testing") two different datasets are generated, which are called "training dataset" and "testing dataset".
The training dataset is generated by splitting up randomly 1/3 of the full dataset (all available data, including development data in order to use an as broad as possible set of training conditions). For generating the "testing dataset", the development data was removed from remaining 2/3 of the data). With this procedure the testing dataset does not contain any development or training data. The testing dataset is, similarly to the full dataset procedure, arranged into seven sub-datasets (one for each study + one combined dataset). To avoid any biasing by the random split of the data, the entire data splitting procedure (between training and testing data) is repeated 2000 times (following Carpenter and Bithel 28), so that 2000 random sampled datasets are generated (bootstrapping), each of them consisting of one combined training sub-dataset and seven sub-datasets for the testing.

Figure 6: Schema of the overall data processing

The metrics were calculated using evalglare [29,30] (versions 1.20 – 2.03), a RADIANCE [31] based tool to evaluate HDR-images. Five different runs with different parameter settings were needed to extract all the information to calculate all metrics investigated for this study (for details, see supplementary material).

3. Results

3.1. Performance of the metrics

In this section the performance of the glare metrics is evaluated. We defined "performance" in section 3 as the:

1. Ability of the metric to correlate with the perception of glare of the subjects, evaluated with the **Spearman-Ranking correlation ρ.**
2. Ability to discriminate between disturbing and non-disturbing glare, evaluated with the Area Under the Curve **(AUC-value)** of the diagnostic test.

Both types of tests are applied to the untrained metrics and therefore do not consider any pre-defined borderline thresholds: they only evaluate if the metric can predict the glare perception in general. The higher the values for both tests - the better the performance. Details about the test can be found in section 2.5.

In Table 4, the Spearman-rank correlation ρ for the glare metrics and different datasets are shown, as well as the ranking of the metrics according to ρ. One can see that the average correlation for fourteen of the metrics are within the "large effect size" ($\rho > 0.5$), while the remaining ones are in the "medium effect size" ($0.3 < \rho \leq 0.5$). Considering all the study datasets separately, only CGI, DGI, DGI_{mod}, DGP and L_{std_win} consistently stay in the "medium effect size" range. All the other metrics fall at least once into the category of "low effect size".

For the second performance evaluation, i.e. the area under the curve (AUC), most of the values lie very close to each other. The average values of the best 15 ranked metrics are between 0.80 and 0.82, which means most of the metrics show an excellent discrimination between disturbing and non-disturbing glare or are very close to it. However, only two metrics (DGP and E_{dir}) are in the acceptable discrimination range for all the investigated datasets. For all the other metrics, the AUC falls for at least one dataset into the "poor" range or even fails (failing for CGI, DGI, DGI_{mod}, PGL, $L_{med\_lowerwin,}$ UGP, UGR and UGR_{exp}).

Another important outcome of the performance evaluation is that on average the metrics perform reasonably well – for both of the applied performance tests. On the other hand, datasets containing slightly more extreme situations like "USA-fabrics" or "DE-gaze" result in very low Spearman correlations or low AUC-values for most metrics, so for these datasets, the performance of metrics falls either into the categories of "low effect size" and/or of "poor discrimination". DGP is the only metrics that falls into the "medium effect size" range and has "acceptable discrimination" for all datasets.

The outcome of the performance evaluation tests does not provide any information about the robustness of the metrics. A high correlation and a high AUC do not automatically imply a high robustness. In case the regression coefficients or borderline-values are changing significantly using different datasets, a metric would not be robust and the applicability would be very restricted, since universal valid thresholds could not be derived. For that reason, the robustness has to be evaluated in addition to the performance.

Table 4: Spearman ranking correlation ρ for the different datasets and metrics. None of the DGP-development-data is used. The dark grey marked values are non-significant. The ranking is based on the average ρ-value. Correlations are within the "large effect size" when $\rho > 0.5$ and in "medium effect size" when $0.3 < \rho \leq 0.5$.

Metric

ρ_{SPEAR} **rank**

0.52

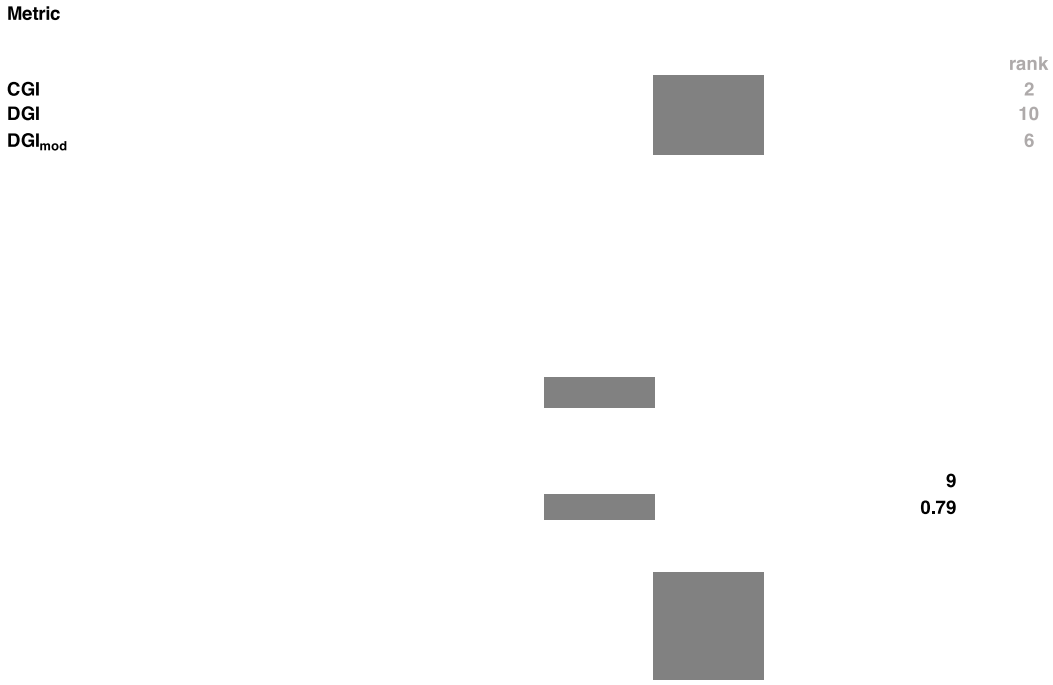
0.57

I propose to highlight large and medium effect size like this. @Jan, I can do it for you, Tilmann

.... values with white text are non-significant. Values in light or dark grey boxes are "medium effect size" or "below medium effect size" respectively.

	Quanta	Gaze	VICE	DEO	Fabric	JP-Office	Combined	average	rank
	ρ_{spear}	ρ_{spear}	ρ_{spear}	ρ_{spear}	ρ_{spear}	ρ_{spear}	ρ_{spear}	ρ_{spear}	
CGI	0,56	0,53	0,67	0,62	0,32	0,38	0,58	0,52	10
DGI	0,5	0,51	0,64	0,56	0,32	0,32	0,54	0,48	15
DGP	0,58	0,54	0,68	0,62	0,58	0,37	0,62	0,57	1
E_v	0,57	0,56	0,67	0,6	*0,26*	0,59	0,55	0,54	4
$L_{40Band,\ avg}$	0,53	0,56	0,55	0,52	0,17	0,57	0,4	0,47	16
L_{avg}	0,57	0,56	0,67	0,68	*0,26*	0,6	0,54	0,55	2

Table 5: Area Under the Curve-values (AUC) for the different datasets and metrics. None of the DGP-development-data is used. For the grey marked cases the AUC values indicate a random behaviour and are thus considered as failing a test. AUC-values ≥ 0.8 or ≥ 0.7 or ≥ 0.6 are considered as excellent, acceptable or poor respectively. AUC < 0.6 is interpreted as fail.



I propose to highlight excellent, acceptable and poor (no fail detected) similar to table 4. @Jan, I can do it for you, Tilmann

3.2. Robustness of the metrics

We defined robustness in 2.5 in general terms as the ability of a metric to deliver meaningful results when applied to different datasets. We investigated this in this study by applying diagnostic tests to the metrics for different datasets, using the same borderline values for all tests and datasets. These

borderline values were derived from training datasets, which are not used for testing the metrics regarding robustness. Two different methods for training and robustness evaluation have been used. Also, the number of failed statistical tests, and the sensitivity with respect to chaning borderline values were treated as a measure for robustness when they are derived from each dataset separately. The applied robustness tests are described in section 2.5.

3.2.1. Analysis of the squared Distance SqD

The squared distance value SqD is a diagnostic test and can be interpreted as both performance and robustness indicator. A small value indicates a high ability to discriminate between "disturbed" and "not disturbed" for a given borderline value. The SqD is a trade-off between the rate of predicting glare situations ("true positive rate TPR") and the rate of predicting non-glare situations ("true negative rate TNR"), giving both the same weight (see 2.5.3.2). The smaller the value the better the metric performs for the used borderline-value. In Table 6 the average SqD values for the different datasets and metrics are shown. The average SqD for each dataset is derived from the 2000 randomly sampled testing-data and uses borderline-values determined by the respective "training dataset" (see also section 2) in order to ensure that no training data is used to evaluate the performance or robustness of the metrics.

A value larger than 0.5 indicates that the data cannot be discriminated reliably. If the SqD varies between the different datasets for a metric, then this indicates a sensitivity to the borderline-value and is a measure for the (non-)robustness of the metric. An example for this can be found in the results of the direct Illuminance E_{dir}. For the datasets DE-Quanta, DE-Gaze, IL-DayVice and AR-DEO, it shows a good performance when looking at the AUC value (between 0.79 and 0.84). However, the SqD value for the AR-DEO dataset (0.39) is significantly higher than for the other three datasets (0.11-0.15). This means that for this dataset, the borderline-value is leading to a poor discrimination, although the metric itself would perform well as indicated by the AUC-value. Therefore, the metric is less robust than others. Following this, the evaluation of the maximum SqD amongst the data-sets within the metrics gives an indication about the robustness of the metric. From all the investigated metrics DGP has the lowest maximum value (0.30), followed by $PGSV_{sat}$ (0.36), E_v (0.36), L_{avg_win} (0.40), E_{dir} (0.41) and L_{avg} (0.42). Only those seven metrics stay below the threshold of 0.5 for all datasets. Furthermore, one can see from Table 6 that DGP, E_v and $PGSV_{sat}$ show the lowest average SqD-values amongst the metrics, which indicates a better performance of these metrics compared to others.

Table 6: Average Squared Distance values of the metrics, derived from the 2000 random-sampled testing- datasets. The cut-off value for each testing-dataset is determined by the respective training-data-set to guarantee that no training data is used to evaluate the performance of the metrics. The dark grey marked cells show values which are larger or equal than 0.5 and indicate therefore that for this case data cannot be discriminated reliably.

Metric	Datasets							Average		Maximum	
	DE-Quanta	DE-Gaze	IL-DayViCE	AR-DEO	US-Fabric	JP-Office	Combined				
	$\overline{SqD}$	$\overline{SqD}$	$\overline{SqD}$	$\overline{SqD}$	$\overline{SqD}$	$\overline{SqD}$	$\overline{SqD}$	$\overline{SqD}$	rank	$\overline{SqD}$	rank
CGI	0.12	0.12	0.11	0.09	0.68	0.46	0.19	0.25	**9**	0.68	**14**
DGI	0.24	0.14	0.13	0.13	0.58	0.46	0.19	0.27	**10**	0.58	**11**
DGI_{mod}	0.17	0.07	0.11	0.10	0.81	0.48	0.20	0.28	**11**	0.81	**19**
DGP	0.10	0.12	0.11	0.09	0.27	0.30	0.15	0.16	**1**	0.30	**1**
E_{dir}	0.10	0.13	0.12	0.39	0.41	0.40	0.20	0.25	**8**	0.41	**5**
E_v	0.11	0.12	0.10	0.12	0.36	0.25	0.17	0.17	**3**	0.36	**3**
GSV	0.13	0.30	0.21	0.63	0.36	0.41	0.26	0.33	**19**	0.63	**12**
L_{40band_avg}	0.17	0.10	0.21	0.25	0.80	0.38	0.30	0.31	**15**	0.80	**17**
L_{avg}	0.10	0.05	0.15	0.20	0.42	0.31	0.16	0.20	**5**	0.42	**6**
L_{avg_win}	0.13	0.14	0.26	0.32	0.40	0.26	0.16	0.24	**7**	0.40	**4**
L_{med}	0.13	0.05	0.39	0.12	0.31	1.00	0.17	0.31	**14**	1.00	**22**
$L_{med\_lowerwin}$	0.21	0.74	0.27	0.80	0.44	0.34	0.24	0.43	**22**	0.80	**18**
L_{med_win}	0.19	0.21	0.25	0.83	0.40	0.28	0.19	0.33	**21**	0.83	**20**
L_{pos_avg}	0.11	0.06	0.12	0.09	0.51	0.31	0.19	0.20	**4**	0.51	**9**
L_{std_win}	0.43	0.13	0.19	0.13	0.99	0.27	0.18	0.33	**20**	0.99	**21**
PGL	0.18	0.10	0.18	0.26	0.66	0.45	0.16	0.29	**12**	0.66	**13**
PGSV	0.09	0.14	0.11	0.28	0.50	0.26	0.19	0.23	**6**	0.50	**7**
$PGSV_{sat}$	0.11	0.12	0.10	0.12	0.36	0.25	0.17	0.17	**2**	0.36	**2**
UGP	0.23	0.14	0.13	0.35	0.73	0.45	0.21	0.32	**17**	0.73	**16**
UGR	0.23	0.14	0.13	0.35	0.73	0.45	0.21	0.32	**16**	0.73	**15**
UGR_{exp}	0.27	0.10	0.15	0.50	0.47	0.50	0.26	0.32	**18**	0.50	**8**
VCP	0.28	0.23	0.21	0.18	0.33	0.53	0.28	0.29	**13**	0.53	**10**
Average	**0.18**	**0.16**	**0.17**	**0.29**	0.52	**0.40**	**0.20**	**0.27**			

3.2.2. Analysis of results for True Positive Rate TPR and True Negative Rate TNR

As mentioned in 2.5.3.3, the interpretation of TPR and TNR has to be done for both at the same time as a high value for one of the two values could be caused by a very low one for the other, originally caused by a non-optimal borderline value. Table 7 gives an overview of the calculated TPR and TNR values for the metrics and the datasets. An example of non-robustness for one dataset is the DGI_{mod}. Whereas it performs well in general, indicated by a high average AUC value (0.81), and reasonably well for most of the datasets looking at the TPR and TNR values, it has a very high TPR (0.96) for the US-Fabric dataset but at the same time a TNR of only 0.10. In that case there would be an over-prediction of 85% of non-glare situations and therefore, for this dataset, the results of the diagnostic test become meaningless even though the AUC value (0.69 for the US-Fabric dataset) indicates a reasonable performance. The reason for this is that the metric is not robust using the same borderline-value for all the datasets, which is an important requirement for the application of a metric. Our evaluations show (see Table 7) that most of the metrics fail such a robustness requirement. The DGP is the only metric not failing this requirement for any of the datasets. Another important result of the TPR and TNR analysis is that the prediction rate (for both TPR and TNR) for five of the metrics (DGP, E_v, L_{pos_avg} and $PGSV_{sat}$) are larger or equal to 0.70. This means that these four metrics have an average prediction-rate of 70-75% for the discrimination between "disturbing glare" and "non-disturbing glare" for the investigated studies. Considering the fact that the metrics are predicting subjective perceptions and that the data were collected with different protocols in different countries and continents, this prediction rate can be considered as reasonably high.

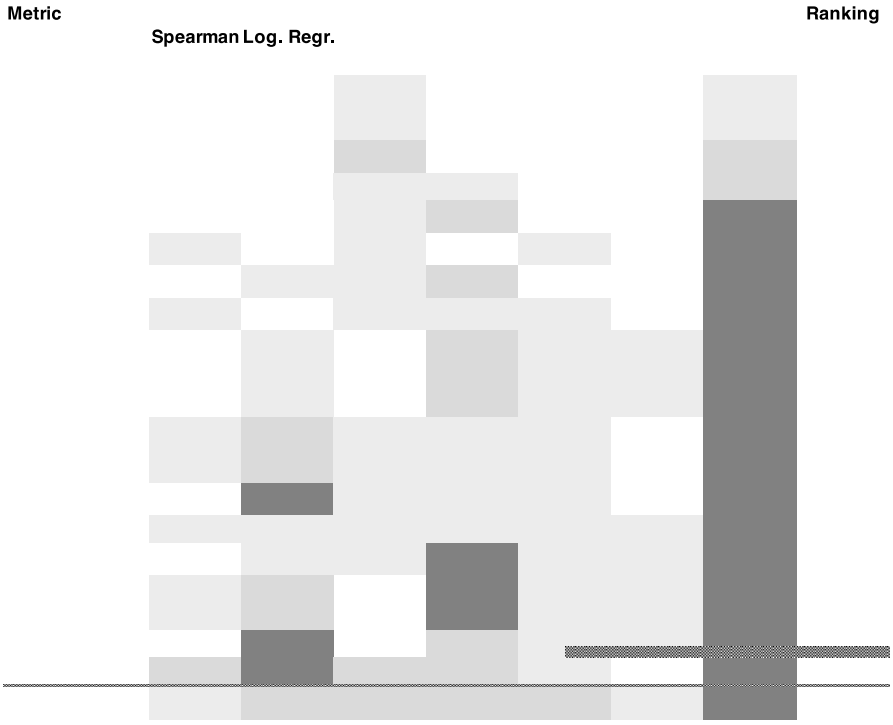
Table 7: Average TPR and TNR values of the metrics, derived from the 2000 random-sampled testing- datasets. The cut-off value for each testing-dataset is determined by the respective training-data-set to guarantee that no training data is used to evaluate the performance of the metrics. The dark grey cells show a failed test where a value less than 0.5 indicates a meaningless prediction rate. The ranking is based on the sum of failed tests.

Metric	Datasets														Minimum		Sum failed	
	DE-Quanta		DE-Gaze		IL-DayViCE		AR-DEO		US-Fabric		JP-Office		Combined					
	$\overline{TPR}$	$\overline{TNR}$	$\overline{TPR}$	$\overline{TNR}$	$\overline{TPR}$	$\overline{TNR}$	$\overline{TPR}$	$\overline{TNR}$	$\overline{TPR}$	$\overline{TNR}$	$\overline{TPR}$	$\overline{TNR}$	$\overline{TPR}$	$\overline{TNR}$	$\overline{TPR}$	$\overline{TNR}$	sum	rank
CGI	0.78	0.78	0.67	0.97	0.75	0.79	0.80	0.80	0.90	0.19	0.73	0.41	0.79	0.64	**0.67**	0.19	**2**	**5**
DGI	0.56	0.83	0.64	0.93	0.76	0.75	0.85	0.70	0.91	0.25	0.95	0.33	0.79	0.62	**0.56**	0.25	**2**	**5**
DGI_{mod}	0.68	0.80	0.78	0.92	0.76	0.77	0.80	0.76	0.95	0.11	0.60	0.45	0.78	0.63	**0.60**	0.11	**2**	**5**
DGP	0.82	0.74	0.67	0.97	0.78	0.77	0.82	0.78	0.70	0.60	0.69	0.58	0.74	0.72	**0.67**	**0.58**	**0**	**1**
E_{dir}	0.86	0.71	0.65	0.97	0.81	0.71	0.91	0.39	0.38	0.85	0.81	0.40	0.69	0.68	**0.38**	**0.39**	**3**	**16**
E_v	0.83	0.72	0.68	0.96	0.80	0.76	0.83	0.72	0.42	0.83	0.76	0.58	0.68	0.75	**0.42**	**0.58**	**1**	**2**
GSV	0.87	0.66	0.49	0.93	0.60	0.77	0.90	0.22	0.43	0.81	0.84	0.38	0.64	0.65	**0.43**	**0.22**	**4**	**20**
L_{40band_avg}	0.71	0.74	0.74	0.88	0.60	0.81	0.76	0.58	0.11	0.95	0.80	0.42	0.54	0.72	**0.11**	**0.42**	**2**	**5**
L_{avg}	0.82	0.74	0.81	0.91	0.85	0.64	0.86	0.59	0.37	0.89	0.47	0.87	0.66	0.79	**0.37**	**0.59**	**2**	**5**
L_{avg_win}	0.68	0.85	0.64	0.95	0.94	0.49	0.94	0.45	0.40	0.84	0.58	0.75	0.67	0.77	**0.40**	**0.45**	**3**	**16**
L_{med}	0.87	0.67	0.82	0.89	0.97	0.38	0.67	0.91	0.59	0.64	0.00	1.00	0.68	0.75	**0.00**	**0.38**	**2**	**5**
$L_{med\_lowerwin}$	0.61	0.80	0.15	0.97	0.95	0.48	0.94	0.12	0.37	0.81	0.73	0.52	0.64	0.67	**0.15**	**0.12**	**4**	**20**
L_{med_win}	0.61	0.82	0.56	0.93	0.96	0.50	0.95	0.09	0.40	0.83	0.54	0.79	0.66	0.74	**0.40**	**0.09**	**2**	**5**
L_{pos_avg}	0.84	0.72	0.78	0.93	0.78	0.73	0.82	0.78	0.30	0.90	0.78	0.50	0.66	0.73	**0.30**	**0.50**	**1**	**2**
L_{std_win}	0.35	0.91	0.66	0.92	0.77	0.66	0.74	0.76	1.00	0.00	0.65	0.70	0.74	0.69	**0.35**	**0.00**	**2**	**5**
PGL	0.60	0.88	0.71	0.92	0.60	0.86	0.91	0.51	0.90	0.19	0.35	0.90	0.68	0.77	**0.35**	**0.19**	**2**	**5**
PGSV	0.79	0.79	0.64	0.96	0.78	0.76	0.87	0.50	0.31	0.90	0.62	0.69	0.63	0.79	**0.31**	**0.50**	**2**	**5**
$PGSV_{sat}$	0.83	0.72	0.68	0.96	0.80	0.76	0.83	0.72	0.43	0.83	0.76	0.58	0.69	0.75	**0.43**	**0.58**	**1**	**2**
UGP	0.57	0.84	0.64	0.94	0.73	0.77	0.89	0.43	0.90	0.16	0.96	0.34	0.79	0.60	**0.57**	**0.16**	**3**	**16**
UGR	0.57	0.84	0.64	0.94	0.73	0.77	0.89	0.43	0.90	0.16	0.96	0.34	0.79	0.60	**0.57**	**0.16**	**3**	**16**
UGR_{exp}	0.90	0.49	0.70	0.96	0.81	0.67	0.91	0.30	0.33	0.87	1.00	0.29	0.72	0.58	**0.33**	**0.29**	**4**	**20**
VCP	0.88	0.50	0.57	0.86	0.57	0.87	0.83	0.63	0.72	0.50	0.54	0.45	0.68	0.60	**0.54**	**0.45**	**2**	**5**
Average	**0.73**	**0.75**	**0.65**	**0.94**	**0.78**	**0.70**	**0.85**	**0.55**	**0.58**	**0.60**	**0.69**	**0.56**	**0.70**	**0.69**	**0.41**	**0.33**		

3.2.3. Failing statistical tests

A similar result for the robustness as for the TPR/TNR evaluation can be derived from the number of failing statistical tests, described in section 2.5.5 (see Table 8). In this table, the failings of all statistical tests are summarized for all the metrics. DGP is the only metric that is not failing any of the applied statistical tests. E_v and $PGSV_{sat}$ are failing only one test; L_{avg} and PGSV are failing two tests. All the other metrics are failing three or more tests, which indicates a very low robustness.

Table 8: Summary table of the number of statistical tests failed for the metrics. The shown number corresponds to the number of datasets for which a statistical test is failed. For the Spearman correlation and the logistic regression, a test treated as failed, if the respective p-value is larger than 0.00227 (Bonferroni-corrected significance level of 0.05 for 22 evaluations, see section 2.5.2). For TPR and TNR, a test is treated as failed if the respective value is less or equal than 0.5. The SqD test is failed, if the value is larger than 0.5. For AUC, a test is treated as failed if the value is less than 0.6.



3.3. Average borderline values

In Table 9 the borderline-values for the investigated metrics are presented as result of the diagnostic tests of the 2000 random-sampled training- datasets. The values distinguish between the four categories of the subjective response scale (details see 2.5.3.4), serve as documentation for this study and help in the interpretation of results from other experiments or simulation results.

Table 9: Average borderline ("cut-off") values of the metrics, derived from diagnostic tests of the 2000 random-sampled

training datasets.

Metric	Imperceptible - Noticeable BIN	Noticeable - Disturbing	Disturbing - Intolerable BDI
CGI	28.7	31.0	34.8
DGI	19.0	19.9	22.4
DGI_{mod}	18.4	19.7	21.9

For most of the metrics a direct comparison of the borderline values of Table 9 with previously published values is difficult since different semantic scales are used (see 2.5.3.4). Therefore, we compare them only for DGP since it was developed with the same semantic scale (Table 10). The calculated borderline values are very close to the originally published ones [41] and the ones used in the European standard EN17037 :

Table 10: Comparison of the average calculated borderline ("cut-off") values of DGP and the originally published values [41] and the values used in the European standard EN17037.

Borderline	Calculated cross-validation study	EN17037 and [41]
BIN	0.34	0.35
BND	0.38	0.40
BDI	0.45	0.45

3.4. Variation of borderline values

If the borderline values are derived from each dataset separately, the variation of the values can serve as a robustness indicator as well. The result of this analysis can be found in Table 11, where we used the normalized RMSE (NRMSE) to quantify the variation of the borderlines between the different datasets using the average borderline values as reference. Five of the metrics show a NRMSE lower than 10% (DGP, DGI_{mod}, DGI, and CGI), which indicates a high robustness. On the opposite side,

seven metrics (L_{std_window}, L_{avg_window}, E_{dir}, L_{med_win}, $L_{med\_lower\_win}$, and PGL) exhibit very high deviations (NRMSE > 50%), indicating a low robustness.

Table 11: Borderline ("cut-off") values of the metrics, derived from the six datasets in comparison with the average borderline value, derived from the training dataset. The normalized RSME is calculated by using the training-data derived value as reference.

	Training-Data	DE-Quanta	DE-Gaze	IL-DayViCE	AR-DEO	US-Fabrics	JP-Office	**NRMSE [%]**	**Ranking**
CGI	**31.0**	30.9	28.5	30.2	37.7	31.6	30.7	**9%**	4
DGI	**19.9**	19.2	18.4	19.4	22.5	22.2	21.1	**8%**	3
DGI_{mod}	**19.7**	19.1	19.4	19.0	22.6	21.9	19.2	**7%**	2

4. Discussion

A quick overview the ranking results of the performance and robustness evaluations of section 3 is shown in Table 12.

The ranking for the average AUC was not considered, since the differences between the various metrics were so small that a ranking could be misleading. Instead we used the results from the average squared distance analysis (together with the Spearman analysis) for the performance ranking. The six highest ranked metrics for both the performance and the robustness evaluation end up being metrics that consider the saturation effect as a main effect in the equation (DGP, $PGSV_{sat}$, E_v, L_{avg}, L_{pos_avg}, PGSV). Metrics based only on contrast or masked areas of the image, as well as purely empirical equations, do not perform as well and are less robust. In the following section, the metrics are discussed in detail.

Table 12: Summary of the rankings of the performance and robustness evaluations of section 3. The metrics are sorted according to the total ranking, which is the average of the overall ranking of performance and robustness.

Metric	total	overall	Spearman

4.1. Metrics based on contrast effect only

Established glare metrics such as CGI, DGI, UGR, VCP as well as modifications of them (DGI_{mod}, UGP) are based only on the contrast effect. In our cross-validation study, the contrast-based metrics perform less well and are also less robust than metrics using the saturation effect in their equation. This might be because all of our studies used workplace positions which were daylight-dominated and exposed to higher light levels (and therefore higher adaptation levels) than during the experiments where these metrics were originally developed. In addition, the exclusion of the saturation effect seems to be an intrinsic disadvantage of these metrics when dealing with large sized glare sources, as is often the case when dealing with daylight. This can be illustrated by the following example, comparing results from a large window façade with a small window façade using white Venetian blinds (images acquired in the DE-Ecco project). For similar conditions (sky type, time of day etc.), increasing the window size to an increase of the overall brightness of the scene (manifested by an increase of the average vertical illuminance in that experiment from 2494 lux to 4468 lux). While the average luminance of the window was almost the same (3032 cd/m^2 for the small window and 2815 cd/m^2 for the large one), 29% of the subjects were disturbed by glare for the small window versus 49% for the large one (see also Table 13). This increase of disturbance-rate is not reproduced by any of the contrast-based metrics, which are calculated using equations where the luminance of the glare source is divided by the background luminance. Since the effect of the increase in size of the glare source (expressed by the solid angle) is not enough to compensate for the

increase of the background luminance, the values of the contrast-based metrics decrease when the window size is increased.

Table 13: Influence of window size on saturation-effect-based and contrast-based metrics.

		Average	Ratio of persons	Saturation effect based metrics		Contrast based metrics		
Window Size	Cases	Window Luminance [cd/m^2]	disturbed by glare	E_v [lux]	DGP [-]	DGI [-]	CGI [-]	UGP [-]
Small	42	3032	29%	2494	0.29	20.5	29.7	0.85
Large	43	2815	49%	4468	0.43	17.8	29.3	0.76

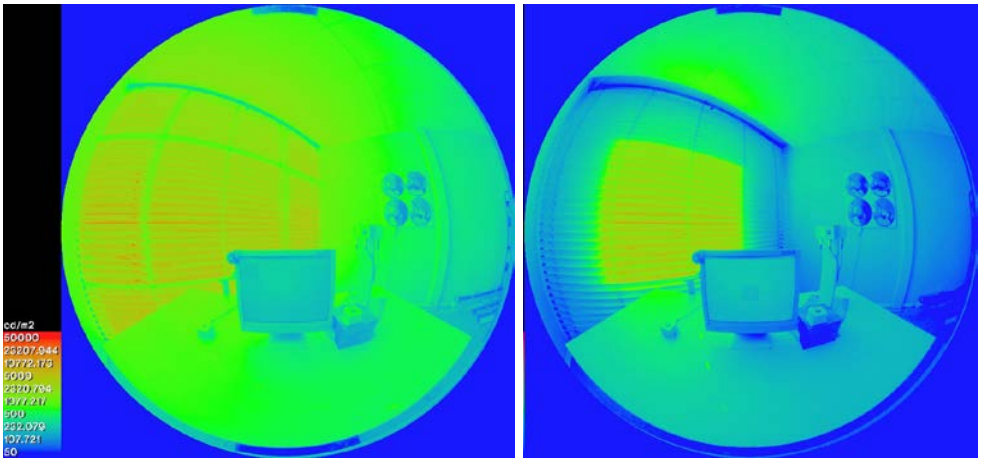


Figure 7: Example images from the DE-Ecco dataset to illustrate the saturation effect. On the left side the large window façade setting is shown, on the right the small window façade setting. Both experimental conditions use the same room and the same white Venetian blind system and the luminance in the window-area is almost the same. For the small window, 29% of the subjects were bothered by disturbing glare whereas for the large one 49%. This increase of disturbing-rate cannot be reproduced by metrics which are only based on the contrast-effect.

In our cross-validation study and amongst the contrast-based metrics, the **CGI** performs best and is also more robust than the other metrics of this category. We assume that this is because the CGI uses the direct illuminance E_{dir} (illuminance induced only by the glare source) to enlarge the effect of the product of $L_s^2 \cdot \omega$.

The **DGI$_{mod}$** performs slightly better than the **DGI**, but the improvement is not significant. The performance of the **UGR** and **UGP** is very similar, because they have the same equation structure. Both of them fail the significance test for the US-fabric dataset and three times the 0.5 threshold of a diagnostic test (for AR-DEO, US-Fabric and JP-Office). Since **UGP** was especially developed for open-plan offices with assumed lower light levels, the low performance in this study can be explained by the aforementioned neglect of the saturation effect for daylight-dominated workplaces. The visual comfort probability **VCP** performs at the lowest end of all the metrics and for five datasets the logistic regression is not significant (in total it fails eight tests on five datasets).

4.2. Metrics based on saturation effect only

Metrics based on the saturation effect use the amount of light at the eye as a main variable in their equation. This effect was first mentioned as "'grand total effect" in [44]. From this category of metrics, in our study we investigated E_v, $PGSV_{sat}$, L_{avg} and L_{avg_pos} and E_{dir}. Except E_{dir} they all performed

reasonably - and better than the metrics based only on the contrast effect. Of the five metrics based purely on the saturation effect, $\mathbf{E_v}$ and $\mathbf{PGSV_{sat}}$ seem to be more robust than the others. E_v and $PGSV_{sat}$ perform quite similarly in all respects. They both fail only one of the tests, and the failures are observed only for the US-Fabrics dataset.
This is to some extent expected due to the presence of the extreme luminance source of the sun in every data point of this particular dataset, the impact of which can never be captured by the total E_v due to the low solid angle of the sun. The intrinsic disadvantage of this category of glare metrics neglecting the contrast effect can be also observed in dim environments. An example for this problematic is shown in following image, where the façade is equipped with a low-transmittance shading system (in that case 2%). This low transmittance leads to a low illuminance (E_v=514 lux), but the sun disk is visible with a luminance of several million cd/m^2. This potential glare source is totally neglected by metrics which are based only on the saturation effect – whereas the contrast-based will show an impact.

The weighting of the luminance by the position index $\mathbf{L_{avg\_pos}}$ does not appear to be an improvement to the non-weighted average luminance $\mathbf{L_{avg}}$ – their results are nearly identical. The direct illuminance $\mathbf{E_{dir}}$, which considers only the glare sources and omits the background luminance for the calculation of its value, performs more poorly than the other metrics based on the saturation effect, and is also slightly less robust (it fails three tests for three datasets).

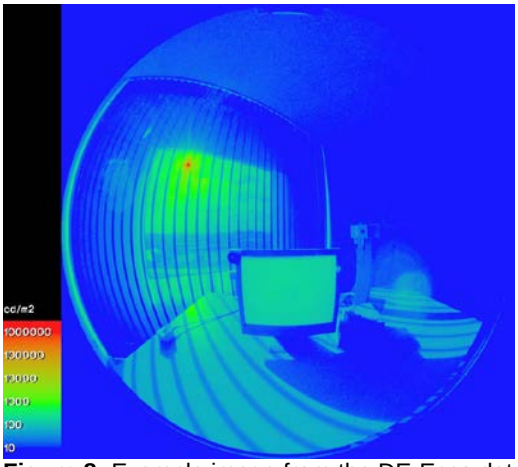


Figure 8: Example image from the DE-Ecco dataset of a situation, where the saturation-effect-based metrics would fail. The image shows a low transmittance shading system, applied as transparent vertical blind behind the glazing.

4.3. Metrics based on both contrast and saturation effects

In this category of metrics, we investigated three metrics: DGP, PGSV and UGR_{exp}. The performance and robustness of these metrics differ significantly, so they will be discussed separately. The **DGP** combines the contrast and saturation effect in an additive manner: it was found to be the most robust and best-performing metric in our study. It is the only metric that did not fail any of the tests. In average across the seven datasets, the Spearman correlation ρ of the DGP is 0.57. The average prediction rate for disturbing glare (True Positive Rate TPR and True Negative Rate TNR)

was about 75%, which means that ¾ of the scenes are predicted correctly and the metric can differentiate between disturbing glare and non-disturbing glare (resp. no glare).
Hirning [4] showed that DGP underestimates glare in open plan office situations, a finding not supported by our study. Given the fact that in open-plan offices, as Hiring was investigating, daylight typically contributes only partly to the total amount of light at the workplace and given the fact of the low sensitivity in the contrast-part in the DGP equation, Hirning's results can be expected. For scenes where overall light levels are rather low and very bright surfaces are visible relatively far from the person (e.g. a window on the other side of the room) and induce a high contrast to the visual task (e.g. computer screen), the DGP might underestimate the perceived glare by the occupants which is the consequence of the limitations of the underlying dataset DGP was developed from. This is also the case for other saturation-effect-driven metrics. However, such scenes were not part of any of the datasets in this cross-validation study: all experiments were conducted with workplace positions close to the window where daylight is the main light source. Future studies should explicitly address these kinds of situations in their design of the experiments.
In Table 14 a recent modification of the daylight glare probability (DGP_{mod}[5]) is compared to the original DGP using Spearman ρ and the AUC. The DGP_{mod} was developed for cases of direct sun visible through shading fabrics. From the results shown in Table 14 no significant improvement can be concluded. Surprisingly the performance of DGP_{mod} for the USA-Fabric dataset is slightly lower than of the original DGP even though this dataset contains only situations with fabric shading devices and was used to develop the metric. The reasons for this should be investigated in a follow-up study.

Table 14: Comparison of the performance (AUC and $\rho_{Spearman}$) for DGP and the modified DGP_{mod}.

AUC	DE-Quanta	DE-Gaze	IL-DayViCE	AR-DEO	US-Fabric	JP-Office	Combined	Average
DGP	0.85	0.91	0.86	0.79	0.72	0.73	0.82	0.81
DGP_{mod}	0.85	0.91	0.85	0.79	0.71	0.75	0.81	0.81
Spearman ρ	**DE-Quanta**	**DE-Gaze**	**IL-DayViCE**	**AR-DEO**	**US-Fabric**	**JP-Office**	**Combined**	**Average**
DGP	0.57	0.54	0.68	0.62	0.37	0.57	0.61	0.57
DGP_{mod}	0.58	0.55	0.68	0.61	0.34	0.59	0.60	0.56

The **PGSV** considers saturation and contrast effects by two separate equations ($PGSV_{con}$ and $PGSV_{sat}$ [15]) which are applied conditionally. The contrast equation is applied only when the ratio between glare source luminance L_s and background luminance L_b is larger than the ratio between average luminance of the scene L_{avg} and the adaptation luminance L_a (which corresponds to the task luminance). For our datasets PGSV is ranked slightly below than $PGSV_{sat}$. This behaviour is surprising - considering the separation between saturation and contrast situations should improve the performance as well as robustness. The lower performance is illustrated in Table 15, where the Spearman correlations and AUC were compared between the three PGSV equations. We conclude from this that the "condition" function to decide between the two equations could be optimized (e.g. needs additional scaling parameters to decide more appropriately) or should be transformed into a summation. Another interesting finding about the PGSV equations is that all of them perform pretty

well for the Japanese dataset, although there were not developed with that data. We hypothesise that the PGSV is adapted to Japanese users. This should be investigated more in depth in a follow-up study.

Table 15: Comparison of the performance (AUC and $\rho_{Spearman}$) for the three PGSV equations.

AUC	DE-Quanta	DE-Gaze	IL-DayViCE	AR-DEO	US-Fabric	JP-Office	Combined	Average
$PGSV_{con}$	0.83	0.85	0.80	0.83	0.68	0.75	0.76	0.79
$PGSV_{sat}$	0.84	0.91	0.86	0.84	0.67	0.75	0.79	0.81
PGSV	0.84	0.89	0.85	0.83	0.68	0.76	0.79	0.81
Spearman ρ	**DE-Quanta**	**DE-Gaze**	**IL-DayViCE**	**AR-DEO**	**US-Fabric**	**JP-Office**	**Combined**	**Average**
$PGSV_{con}$	0.56	0.46	0.58	0.59	0.28	0.59	0.51	0.51
$PGSV_{sat}$	0.57	0.56	0.67	0.60	0.26	0.59	0.55	0.54
PGSV	0.57	0.51	0.66	0.54	0.28	0.60	0.54	0.53

The $\mathbf{UGR_{exp}}$ also combines additively the saturation and contrast effects. Unlike the DGP, it uses the logarithm of the average luminance for the saturation effect and uses the glare source luminance only by the power of 1 in the contrast term [10]. UGR_{exp} performs at the lower end of the investigated metrics and also seems to be not robust (six failings in four datasets). We assume that this behaviour is caused by the smaller influence of the glare sources, which are proportional to $\log(L_s/L_b*\omega/P^2)$ while typical glare equations are proportional to $\log(L_s^2/L_b*\omega/P^2)$. In [12] it was shown that the logarithmic function applied to E_v ends up in a lower correlation to the ratio of people disturbed by glare than the linear $E_{v,}$. We thus conclude that the logarithmic function applied to the average luminance L_{avg} is disadvantageous.

4.4. Equations based on the contrast effect and absolute thresholds

Two recently published equations are based on the summation of the contrast between glare source and task luminance L_s:L_t and the glare source luminance. The **PGL** considers neither the size of a glare source nor the saturation effect in its equation and uses the contrast effect in a linear approach. This may be why this equation is not robust (failing six tests for three datasets) and performs poorly compared to the other metrics.

The **GSV** also uses the contrast as a linear function, but it uses as absolute threshold the area-ratio of luminance-values larger than 2000 cd/m^2 in Guth's field of view. The size of the glare source is thus accounted for, but not its luminance (because it is just an area ratio). We conclude that the linear contrast approach and the missing luminance value in the absolute-threshold term leads to the weak robustness (failing nine tests on four datasets) and low performance. However, one should note that GSV is addressing in particular sun-spots in or close to the task area, though it was applied to the entire dataset, which consists in large fraction of very different lighting scenes. Therefore, further investigation would be needed before drawing any firm conclusion.

4.5. Empirical equations

In this section we discuss empirically derived equations, which neither use the saturation nor the contrast effect. $\mathbf{L_{40band\_avg}}$ uses only average luminance from a band of 40° around a horizontal axis of the image. Obviously, any glare source outside this band is not considered by this metric. Such a

situation is shown Figure 9. We assume that this is why this metric performs rather poorly and is also not as robust (5 failings for 3 datasets) as many other metrics.

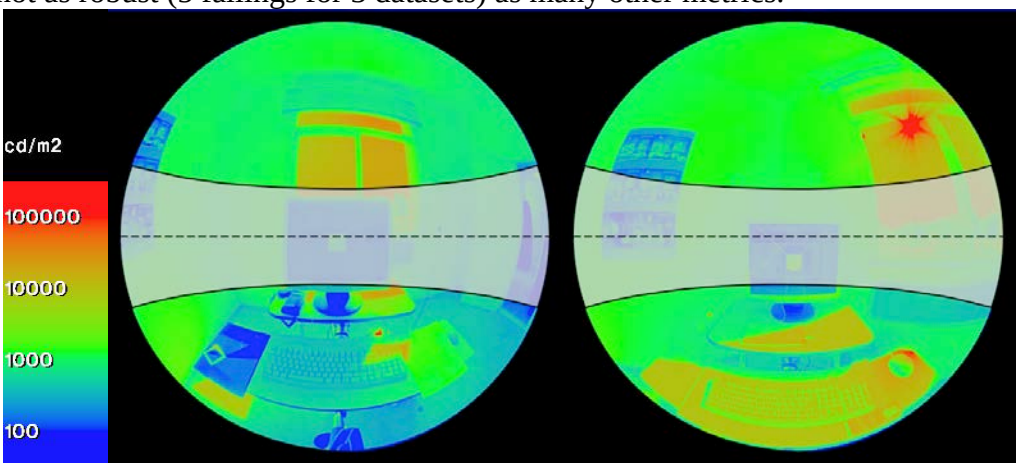


Figure 9: Example images of situations, where the 40° band would miss potential glare sources, taken from the IL-DayViCE dataset.

The average luminance of the window, $\mathbf{L_{avg\_win}}$, is less sensitive to this afore mentioned problem since the main glare source in daylit scenes can typically be found in the window. We assume that this is why it performs better. However, this metric does not account for either the contrast effect or for different window sizes (respectively any solid angle of the window). Also, it neglects reflections on internal surfaces or disturbing sun-patches inside the room. We assume this is why the metric is less robust than others and a universally-applicable borderline-value is not reliable. Consequently, it would treat both scenes shown in Figure 7 the same.

The three metrics that use a median-value of the image or parts of the image (L_{med}, $L_{med\_lowerwin}$ and L_{med_win}) as well as the standard deviation of the luminance of the window L_{std_win} were derived empirically and have nothing in common to accepted influence-factors of glare perception (e.g. luminance of glare source, solid angle of glare source)[45]. However, the median luminance of the image $\mathbf{L_{med}}$ performs reasonably well in our study and fails only three tests for two datasets. An intrinsic problem of all median-based metrics is that they do not account per se for the high luminances in the image, despite that it is common knowledge that the high luminances in the field of view have a major influence on glare perception. This is a problem for a very inhomogeneous scene (e.g. with the sun disk visible behind a screen), since this potential glare risk will not be addressed by any median value. Even more critical is the situation for median-based metrics using only parts of the image like the two other investigated metrics $\mathbf{L_{med\_lowerwin}}$ and $\mathbf{L_{med\_win}}$ since the masking might miss potential glare sources when calculating the median value. We assume that this is also the reason, why these two metrics perform worse and less robust than L_{med}. The results show that the smaller the remaining area of the masking is, the lower is the performance of the median-based metric.

The intrinsic disadvantage of purely empirical derived glare metrics is the neglect of perception mechanisms in their equation. This causes a large uncertainty when the lighting conditions differ significantly to the conditions of development. Therefore, we cannot recommend to use them for glare analysis. We did not evaluate the overall visual discomfort or satisfaction in our analysis and

therefore we cannot make conclusions for these kinds of evaluations. Overall visual discomfort, which these metrics are aiming for [3], include more influencing factors such as view, light levels, inhomogeneity, colour and glare is therefore only one out of several variables influencing the overall visual perception.

4.6. Stimuli range bias

Fotios [23] pointed out that *stimuli range bias* is a common problem in glare research. The low variation of the borderlines (see 3.4) in our study means we cannot observe a stimuli range bias, although we use data based on different stimuli ranges, setups, research groups and climatic/cultural conditions. We assume that the applied 4-point Likert scale [26] is less sensitive to get biased than for example, a linear scale from 0-10. We also assume that a more realistic (and unbiased) choice of the subjects on the rating scale was possible, because all the experiments were executed in office like test rooms, where the subjects can relate to their normal working environment easily. Therefore, we conclude that the application of the 4-point scale (imperceptible, noticeable, disturbing and intolerable) in combination with the use of office-like test-room can avoid a stimuli range bias.

4.7. Limitations

Experimental setup

Although the data were acquired from several different groups in different countries and continents and are therefore very broad, the experimental setup was restricted in all participating studies to a daylight-dominated workplace position, similar to a small office configuration where people sit close to the window. Therefore, the results cannot be extrapolated per se to lighting scenarios which differ significantly (e.g. dim and large open-plan spaces equipped with solar control glazing or spaces mainly lit by rooflights).

Working environment

In addition, all experiments were conducted in controlled environments, where the subjects were invited to participate to the experiment. Therefore, the subjects were exposed to a new environment, which might differ to their normal workplace, although all setups tried to simulate a real office as close as possible. This exposure to another environment might lead to a different perception and acceptance than if the experiments had been conducted in a real environment. However, the advantages having a controlled environment, where the experimental conditions can be set according to the research question and potentially influencing factors can be kept constant, compensates the afore mentioned disadvantage by far.

Glare source detection

For that study we applied the task-driven glare source detection algorithm (see 1.1 and supplementary material), which is supposed to be the most robust and effective method as of yet [46]. However, not all the images were checked, if the detection algorithm was indeed the “best” for the specific scene. The authors randomly checked images for reliable glare source detection. But it must be said that there is no commonly accepted rule to define a glare source in an image and this check of correct glare source detection relies on experience and intuition of the researchers. In general a change of detection parameters can lead to different results[47]. However, we assume that a change in

parameter settings for scenes, where the parameters were not optimal, will not change the overall outcome of the study since there might be only few cases where this would have to be applied.

5. Conclusion and Outlook

The results of this cross-validation study show that metrics that consider the saturation-effect as a main glare effect in their equation perform better and more robustly than purely contrast-based metrics or purely empirical derived metrics. This outcome is valid for daylight-dominated workplaces, though the results might not be 100% transferrable to scenarios which differ significantly (e.g. open-plan offices with overall low light levels).

Spearman correlations in the range of 0.55-0.60 as well as average prediction rates to distinguish between disturbing and non-disturbing glare of 70-75% for several metrics show that their results are trustable. Therefore, a poor performance of glare metrics cannot be observed from this cross-validation study.

In this study, DGP performed best amongst the tested metrics and was found to be the most robust one, since it was the only one not failing any of the applied tests for any dataset. Amongst the contrast-based metrics, the CGI performs best and is also more robust than the other metrics of this category and might be a good choice for evaluating scenes where it is known that saturation does not play any role (e.g open-plan offices with low daylight contribution), though this assumption would have to be confirmed by experiments. The purely empirical derived glare metrics like $L_{med\_lowerwin}$, L_{med_win} or L_{40band_avg} were found not to be robust against lighting conditions which differ significantly to their developing lighting conditions and cannot be recommended to be used for glare analysis.

Overall it remains a challenge to have a glare metric perform reliably in all possible lighting scenarios, architectural or cultural contexts. It is expected that mainly saturation-effect-driven metrics like DGP will perform poorly in dim lighting environments, as it was shown by Hirning [4]. Therefore, upcoming research studies should aim to optimize the combination of contrast-driven and saturation-effect-driven terms in the metric's equations, for example in DGP, PGSV or UGR_{exp}, as their structure already includes these terms. Also, it can be expected that an inclusion of other influencing factors[45] (like culture, contrast sensitivity etc.) in the equations of the glare metrics are likely to improve their performance and robustness.

Declaration of conflicts of interests

The author(s) declared no potential conflicts of interests with respect to the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to thank Clotilde Pierson for the fruitful scientific discussions on the paper and her review and comments on it. The authors also thank Christoph Reinhart for his comments.

Funding

The Ecole Polytechnique Fédéral de Lausanne (EPFL) and the Swiss national Science Foundation SNF (contract no. 205121_157069) supported this cross-evaluation study and the work for the

underlying DE-Gaze dataset. The DE-Ecco and DE-Quanta study was supported by the Fraunhofer-Institute for Solar Energy Systems ISE and the Deutsche Forschungsgemeinschaft DFG (WA1155/5-1). The European Commission supported under the contract ENK6-CT-2002-00656 the Ecco-Build project for the underlying data DE-Ecco and DK-Ecco.
The ISR-DayViCE study was supported by funding from the Israel Ministry of Energy Water and National Infrastructure under contract 2006-8-44/26-11-013.
The work for JP-office was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) (No.21360281).
The USA-Fabrics study was supported by the Alcoa Foundation and Lutron Electronics Co Inc
The ARG-DEO study was funded by National Council of Scientific and Technical Research (Consejo Nacional de Investigaciones Científicas y Técnicas – CONICET) and National Agency of Scientific and Technological Research (Agencia Nacional de Promoción Científica y Tecnológica – ANPCyT) under contracts PICT 2013-2089 and PIP 2009 - A Pattini – CONICET.
The ARG-DEO study was supported by Institute of Environment, Habitat and Energy (Instituto de Ambiente, Hábitat y Energía - INAHE) CCT CONICET Mendoza, Argentina.

References

1. Kuhn TE. State of the art of advanced solar control devices for buildings. *Solar Energy* 2017; 154: 112–133.

2. Van Den Wymelenberg K, Inanici M. A Critical Investigation of Common Lighting Design Metrics for Predicting Human Visual Comfort in Offices with Daylight. *LEUKOS* 2014; 10: 145–164.

3. Van Den Wymelenberg K, Inanici M. Evaluating a New Suite of Luminance-Based Design Metrics for Predicting Human Visual Comfort in Offices with Daylight. *The Journal of the Illuminating Engineering Society* 2015; 12: 113–138.

4. Hirning MB, Isoardi GL, Cowling I. Discomfort glare in open plan green buildings. *Energy and Buildings* 2014; 70: 427–440.

5. Konstantzos I, Tzempelikos A. Daylight glare evaluation with the sun in the field of view through window shades. *Building and Environment* 2017; 113: 65–77.

6. Einhorn HD. Discomfort glare: a formula to bridge differences. *Lighting Research & Technology* 1979; 11: 90–94.

7. International Commission On Illumination (CIE). *Discomfort Glare in Interior Lighting*. 117:1995, 1995.

8. Hopkinson RG. Glare from daylighting in buildings. *Applied ergonomics* 1972; 3: 206–215.

9. Chauvel P, Collins J, Dogniaux R, et al. Glare from windows: current views of the problem. *Lighting research and Technology* 1982; 14: 31–46.

10. Fisekis K, Davies M. Prediction of discomfort glare from windows. *Lighting Research and Technology* 2003; 35: 360–371.

11. Wienold J, Christoffersen J. Evaluation methods and development of a new glare prediction model for daylight environments with the use of CCD cameras. *Energy and Buildings* 2006; 38: 743–757.

12. Wienold J. *Daylight glare in offices*. Fraunhofer-Verlag, Stuttgart; ISBN: 978-3-8396-0162-4, http://publica.fraunhofer.de/documents/N-141457.html (2010).

13. Suk JY, Schiler M, Kensek K. Absolute glare factor and relative glare factor based metric: Predicting and quantifying levels of daylight glare in office space. *Energy and Buildings* 2016; 130: 8–19.

14. Velds M. *Assessment of lighting quality in office rooms with daylighting systems*. Delft University of Technology, 2000.

15. Iwata T, Osterhaus W. Assessment of Discomfort Glare in Daylit Offices Using Luminance Distribution Images. Vienna, Austria, 2010, pp. 174–179.

16. Tokura M, Iwata T, Shukuya M. Experimental study on discomfort glare caused by windows, part 3: Development of a method for evaluating discomfort glare from a large light source. *Journal of Architectural Planning Environment Engineering* 1996; 17–25.

17. Garretón JAY, Colombo EM, Pattini AE. A Global Evaluation of Discomfort Glare Metrics in Real Office Spaces with Presence of Direct Sunlight. *Energy and Buildings*. Epub ahead of print 2018. DOI: https://doi.org/10.1016/j.enbuild.2018.01.024.

18. Rea MS. *The IESNA lighting handbook: reference & application*. New York, NY: Illuminating Engineering Society of North America, 2000.

19. Yamin Garretón JA, Rodriguez RG, Ruiz A, et al. Degree of eye opening: A new discomfort glare indicator. *Building and Environment* 2015; 88: 142–150.

20. Sarey Khanie M, Stoll J, Einhäuser W, et al. Gaze and discomfort glare, Part 1: Development of a gaze-driven photometry. *Lighting Research and Technology*; published online before print. Epub ahead of print 30 June 2016. DOI: 10.1177/1477153516649016.

21. Moosmann C, Wienold J, Wagner A, et al. Ermittlung relevanter Einflussgrößen auf die subjektive Bewertung von Tageslicht zur Bewertung des visuellen Komforts in Büroräumen.

Abschlussbericht, https://publikationen.bibliothek.kit.edu/1000034968 (2012, accessed 2 May 2018).

22. Erell E, Kaftan E, Garb Y. Daylighting for Visual Comfort and Energy Conservation in Offices in Sunny Regions. In: *Proceedings of the 30th PLEA International Conference – Sustainable Habitat for Developing Societies*. Ahmedabad, 2014.

23. Fotios S. Correspondence: New methods for the evaluation of discomfort glare. *Lighting Research & Technology* 2018; 50: 489–491.

24. Fotios S. Research Note: Uncertainty in subjective evaluation of discomfort glare. *Lighting Research & Technology* 2015; 47: 379–383.

25. Kent MG, Fotios S, Altomonte S. Order effects when using Hopkinson's multiple criterion scale of discomfort due to glare. *Building and Environment* 2018; 136: 54–61.

26. Osterhaus WKE, Bailey IL. Large Area Glare Sources and Their Effect on Discomfort and Visual Performance at computer Workstations. *Conference Record of the 1992 Ieee Industry Applications Society Annual Meeting, Vols 1 and 2* 1992; 1825–1829.

27. Christoffersen J, Wienold J. *Assessment of user reaction to glare: All systems*. ECCO-SBI-0406-01, Danish Building Research Institute SBI, http://vbn.aau.dk/en/publications/assessment-of-user-reaction-to-glare-report-eccosbi040601(7d0cd480-8211-11dc-b593-000ea68e967b).html (2006, accessed 7 June 2018).

28. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 2000; 19: 1141–1164.

29. Wienold J, Christoffersen J, Reetz C, et al. Evalglare-A new RADIANCE-based tool to evaluate daylight glare in office spaces. *3rd International Radiance Workshop*, http://www.sbi.dk/indeklima/lys/belysning-i-kontorer/evalglare-a-new-radiance-based-tool-to-evaluate-daylight-glare-in-office-spaces (2004, accessed 19 November 2015).

30. Wienold J, Andersen M. Evalglare 2.0: New features, faster and more robust HDR-image evaluation. In: *3rd International Radiance Workshop*. Padova, Italy, https://www.radiance-online.org/community/workshops/2016-padua/presentations/211-Wienold-Evalgaare2.0.pdf (2016).

31. Ward-Larson G, Shakespeare R. *Rendering with radiance: the art and science of lighting visualization*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., http://portal.acm.org/citation.cfm?id=286090 (1998).

32. Siegel S. *Nonparametric Statistics for the Behavioral Sciences*. 1st edition. New-York: McGraw-Hill, 1956.

33. Rodriguez RG, Yamín Garretón JA, Pattini AE. An epidemiological approach to daylight discomfort glare. *Building and Environment* 2017; 113: 39–48.

34. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press, 1988.

35. Ferguson CJ. An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice* 2009; 40: 532–538.

36. David W. Hosmer, Stanley Lemeshow. *Applied Logistic Regression*. Wiley-Blackwell. Epub ahead of print 2005. DOI: 10.1002/0471722146.index.

37. Coolican H. *Research Methods and Statistics in Psychology*. 6th ed. Psychology Press, https://books.google.ch/books?hl=it&lr=&id=8DJFAwAAQBAJ&oi=fnd&pg=PP1&dq=hugh+coolican&ots=BSeeshI8eW&sig=--W_waAGHMrmDl4VsNsd6ZuOFaA (2014).

38. Safari S, Baratloo A, Elfil M, et al. Evidence Based Emergency Medicine; Part 5 Receiver Operating Curve and Area under the Curve. *Emerg (Tehran)* 2016; 4: 111–113.

39. Luckiesh M, Guth SK. Brightness in Visual Field at Borderline Between Comfort and Discomfort (BCD). *Illuminating engineering*.

40. Jakubiec J, Reinhart C. The 'adaptive zone' - A concept for assessing discomfort glare throughout daylit spaces. *Lighting Research and Technology* 2012; 44: 149–170.

41. Reinhart CF, Wienold J. The daylighting dashboard – A simulation-based design analysis for daylit spaces. *Building and Environment* 2011; 46: 386–396.

42. Bellia L, Cesarano A, Iuliano GF, et al. Daylight Glare: A review of discomfort indexes. In: *Proceedings of Visual Quality and Energy Efficiency in Indoor Lighting: Today for Tomorrow*. Rome, Italy, 2008.

43. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochem Med (Zagreb)* 2016; 26: 297–307.

44. Iwata T, Tokura M. Examination of the limitations of predicted glare sensation vote (PGSV) as a glare index for a large source: Towards a comprehensive development of discomfort glare evaluation. *International Journal of Lighting Research and Technology* 1998; 30: 81–88.

45. Pierson C, Wienold J, Bodart M. Review of Factors Influencing Discomfort Glare Perception from Daylight. *LEUKOS* 2018; 0: 1–38.

46. Pierson C, Wienold J, Bodart M. Daylight Discomfort Glare Evaluation with Evalglare: Influence of Parameters and Methods on the Accuracy of Discomfort Glare Prediction. *Buildings* 2018; 8: 94.

47. Khanie MS, Wienold J, Andersen M. A sensitivity analysis on glare detection parameters (remote presentation). In: *13th International Radiance Workshop 2014*. 2014.