Power analysis, sample size and assessment of statistical assumptions - improving the evidential value of lighting research

J. Uttley[1]

[1] School of Architecture, University of Sheffield, UNITED KINGDOM
j.uttley@sheffield.ac.uk

Power analysis, sample size and assessment of statistical assumptions - improving the evidential value of lighting research

Abstract

The reporting of accurate and appropriate conclusions is an essential aspect of scientific research, and failure in this endeavour can threaten the progress of cumulative knowledge. This is highlighted by the current reproducibility crisis, and this crisis disproportionately affects fields that use behavioural research methods, as in much lighting research. A sample of general and topic-specific lighting research papers were reviewed for information about sample sizes and statistical reporting. This highlighted that lighting research is generally underpowered and, given median sample sizes, is unlikely to be able to reveal small effects. Lighting research most commonly uses parametric statistical tests, but assessment of test assumptions is rarely carried out. This risks the inappropriate use of statistical tests, potentially leading to Type I and Type II errors. Lighting research papers also rarely report measures of effect size, and this can hamper cumulative science and power analyses required to determine appropriate sample sizes for future research studies. Addressing the issues raised in this paper related to sample sizes, statistical test assumptions, and reporting of effect sizes, can improve the evidential value of lighting research.

Keywords: Sample size; power analysis; effect size; statistical test; Type I II errors

1. Introduction

"In the fields of observation chance favours only the prepared mind"

- Louis Pasteur, 7 December 1854

'Eureka' moments are not frequent in science and the scientific endeavour is characterised by the gradual accumulation of knowledge through empirical methods. This relies on evidence that is reliable. As a minimum, evidence should be reported in a manner that allows external verification of its veracity. This allows the reader to judge how appropriate the research conclusions are. One of the cornerstones of science that aims to support this external verification is the peer-review process. This review of research work by experts is designed to filter out poor-quality and unreliable research findings. Peer review has its limitations [Jefferson et al, 2002; Ware, 2008], and may not have been successful in many scientific fields in ensuring the quality of published research, as a large number of published research findings may be false [Ioannidis, 2005]. Publication bias means the vast majority of published findings are positive and support the research hypothesis, and do not provide a representative sample of all scientific studies carried out [Sterling, Rosenbaum & Weinkam, 1995]. This is a problem that is particularly prevalent for human factors research in lighting, as psychological and behavioural science has the highest proportion of studies reporting positive results compared with other scientific disciplines [Fanelli, 2010]. Publication bias may help explain the current reproducibility crisis affecting many sciences but particularly psychological and behavioural science. The Open Collaboration Project recently attempted replications of 100 studies published in three major psychology journals in 2008. Ninety seven percent of the original studies reported significant findings, compared with only 36% of the replication studies. Mean effect sizes in the replications were also half the magnitude of those found in the original studies.

At the heart of publication bias and the reproducibility crisis is the occurrence of Type I errors (false positive findings) and Type II errors (false negative findings). We use statistical methods in science in an attempt to avoid making claims that in reality may be a Type I or Type II error. Null Hypothesis Statistical Testing (NHST, Hubbard & Ryan, 2000) produces a p-value that represents the probability of obtaining the result (or something more extreme) assuming there was no real effect or difference between the groups or measures being tested (the 'null' hypothesis). The p-value does not explicitly refer to the probability of the null hypothesis being true, but it does provide a "measure of the strength of evidence against H_0 [the null hypothesis]" [Dorey, 2010, p. 2297]. Abelson referred to "discrediting the null hypothesis" based on the p-value from a statistical test [Abelson, 1995, p.10]. A smaller p-value provides stronger evidence against the null hypothesis. By convention, in the field of lighting research and most other scientific disciplines, we use a threshold of $p < .05$ to indicate a significant or 'real' effect, based on proposals by Fisher [1925]. However, Fisher himself recognised this threshold was arbitrary and debate is ongoing about its use. The reproducibility crisis has led some researchers to suggest a stricter threshold of .005 should be used [Benjamin et al, 2017], to reduce the number of Type I errors reported in the scientific literature. Other researchers suggest this is unwise, and instead we should justify all experimental design and analytical choices made, including the p-value threshold used to identify a real effect [Lakens et al, 2017].

The debate over p-value thresholds and their use, the existence of publication bias, and the reproducibility crisis all raise questions regarding the evidential value within scientific research in general and within lighting research specifically. A significant consequence of incorrect conclusions made within the research literature is the promulgation of theoretical concepts or methods without appropriate evidence. There are a number of examples of this within lighting research. Veitch [2001] highlighted the example of guidelines for the lighting of internal spaces [Rea & IESNA, 1993; see DiLaura et al, 2011, for newest edition] being based on evidence from one unreplicated study with methodological limitations [Flynn et al, 1979]. Work by Kruithof published in 1941 [Kruithof, 1941] identified combinations of illuminance and correlated colour temperature that supposedly produced pleasing visual conditions for interior lighting. These results have been widely used to support lighting design rules and practice, despite evidence against Kruithof's results [e.g. Boyce & Cuttle, 1990; Davis & Ginthner, 1990; see Fotios, 2017, for a review]. Fotios and Goodman [2012] also highlighted how current guidelines for pedestrian road lighting are based on flawed interpretation of a single unreplicated study by Simons et al [1987].

The evidential value of a study and its contribution to cumulative scientific progress relies on appropriate experimental design and statistical reporting. A critical consideration for any experiment is the sample size used and the experiment's ability to avoid making a Type I or Type II error. The average power of experiments in a range of fields, including cognitive neuroscience, biomedical sciences and ecology [Button et al, 2013; Dumas-Mallet et al, 2017; Lemoine et al, 2016], is low. It is currently not known whether this is the case also in the lighting field of research. Another factor that may negatively impact on the evidential value of a study is the inappropriate use and reporting of statistical tests. Previous research has identified frequent inconsistencies in the reporting of statistics [Bakker & Wicherts, 2011; Garcia-Berthou & Alcaraz, 2004; Nuijten et al, 2016]. However it is not just incorrect statistical reporting that jeopardises the evidential value of a study - the appropriate use of statistical tests in the first place is an important consideration [Thiese, Arnold & Walker, 2015]. Even when statistical tests may be used and reported correctly, and are based on an appropriately-powered experimental design, evidential value can still be limited when information about the size of an effect found in a study is not reported. The reporting of only p-values is not sufficient to convey valuable information about the effect being investigated [Rothman, 2014]. Reporting of effect sizes increases the information content within a paper and facilitates the inclusion of its results into a wider synthesis of evidence, e.g. through meta-analysis.

To assess the inappropriate use of statistical tests, effect sizes and their reporting, and sample sizes and power within lighting research, a sample of lighting research papers were reviewed. Implications of the findings from this review are outlined alongside discussions about how to improve the evidential value of lighting research.

2. Review of statistical reporting within lighting publications

The review examined a sample of general research papers related to lighting, and a sample of research papers related to a specific topic within lighting. This method of using two different types of samples provides both a 'broad but shallow' and a 'deep but narrow' selection of papers. In addition, variations in statistical practices may exist between research

areas within lighting, and this dual-sampling approach allows us to confirm whether statistical practices across a generalised sample of lighting papers represent those used within a specific field. For the sample of general, cross-topic lighting papers, those published in LEUKOS and Lighting Research & Technology during 2017 were included in the review. These two journals are the most prominent outlets for lighting-specific research. For papers about a specific lighting topic, those related to the subject of spatial brightness were selected. There have been a significant number of papers published on this topic, as highlighted by the review carried out by Fotios et al [2015], which identified 70 studies of spatial brightness. Only papers included in Fotios et al's review and published since 2000 were assessed, to better reflect more recent research practice.

For both the general lighting papers published in 2017, and the spatial brightness papers published since 2000, only those that involved research with human participants were included in this review. Basic information was recorded about the sample size, research design (between-subjects, within-subjects or both), statistical tests used, checks of assumptions used in statistical tests, and reporting of effect sizes. A summary of this information is given in Table 1. This table highlights that ANOVAs and related tests (e.g. F-test, MANOVA) are the most frequent type of statistical test used in the papers included in the review, supporting findings that ANOVA is the most common test used in other areas of research such as social psychology [e.g. Kashy et al, 2009]. Other parametric statistical tests such as correlation, regression and t-tests were also found to be in common use in lighting research papers. Parametric tests rely on certain assumptions about the way data were collected and the way they are distributed (discussed later in this paper), yet the review found that test assumptions were rarely assessed.

Table 1 also highlights the median sample sizes used in studies included in the review, for within-subjects and between-subjects designs. The sample size has a major influence on the sensitivity of a study and its ability to reveal something real about the population that has been sampled. A sample that is too small will be unable to reveal a real effect (resulting in a Type II error). Using a larger sample may be a waste of resources however, if a smaller sample would be adequate to reveal the effect under investigation. Selection of sample size is therefore a critical experiment design choice, yet almost all of the studies that were assessed in the review failed to justify the sample size used. This included a lack of reporting of any preliminary power analysis carried out to justify sampling decisions, and a lack of discussion about the size of effect that could be revealed or the size of effect that was anticipated. In addition, a low proportion of studies reported any type of effect size measure, and this was particularly the case for reporting of group differences (the majority of effect size measures that were reported were R^2 values from a regression).

TABLE 1. Summary of basic findings from review of recent lighting papers, and papers relating to spatial brightness research topic since 2000.

Variable	**LR&T and LEUKOS journal 2017**	**Spatial brightness papers since 2000**
Total number of	83	N/A

journal papers in 2017		
Number of papers included in final review	37	13
Research design	84% (31) within-subjects 11% (4) between-subjects 5% (2) mixed (within- & between-subjects)	54% (7) within-subjects 15% (2) between-subjects 31% (4) mixed (within- & between-subjects)
Median sample size	23 for within-subjects 30 for individual groups in between-subjects*	35 for within-subjects 21 for individual groups in between-subjects*
Statistical tests used	ANOVA (including MANOVA) = 62% T-test = 22% Regression = 27% Correlation = 32% Wilcoxon signed-rank = 11% Kruskal-Wallis / Friedman test = 11% Other** = 19% No inferential statistics reported = 8%	ANOVA (including MANOVA) = 46% T-test = 23% Regression = 8% Wilcoxon signed-rank / Friedman test / Kruskal-Wallis = 8% Other** = 23% No inferential statistics reported = 23%
Assessment of assumptions of statistical test(s) used	24% of papers (9)	15% of papers (2)
Report measure of effect size	30% of papers (11)	8% of papers (1)

* In studies with unequal group sizes, the mean group sample size was used in the calculation of the median group size across all studies

** Includes McNemar test, Cochran's Q, post hoc Tukey tests, Standardised Residual Sum of Squares, variance stable rank sums, binomial test, Dunn Rankin test

One further conclusion that emerged from the review of statistical reporting in the selected papers was the variation in exactly what statistics are reported when inferential tests are used. These differences included whether measures of variation such as standard deviation were reported, whether the actual test statistic and associated degrees of freedom were reported, and the precision with which p-values were reported, particularly when a test was not significant. In such cases, p-values were frequently not reported at all.

This review of a sample of general and topic-specific lighting research papers highlights three issues that may compromise the evidential value of research literature within the lighting field. The first is the widespread but potentially inappropriate use of parametric statistical testing, given that only a minority of studies confirm that test assumptions have

been assessed and met. The second is the failure of lighting papers to report measures of effect sizes. The third issue is the relatively small sample sizes used in experiments.

3. Assessment of assumptions required by parametric tests

3.1 Statistical test assumptions

The review of lighting papers described in Section 2 highlighted that parametric tests are the most common type of statistical test used. As the name implies, parametric statistical tests are based on the assumption of certain parameters about the data being tested and the conditions in which it was obtained. However, the review indicated that only 22% of the 50 papers examined actually reported assessing assumptions related to the use of statistical tests. This is concerning as violations of these assumptions can lead to invalid or inappropriate conclusions based on the results of the test and we "stop being able to draw accurate conclusions about reality" [Field, Miles & Field, 2012, p. 167], although the magnitude of the violation will influence the extent to which the conclusions of the test can be accepted. Most parametric methods, including those most commonly used in lighting research such as ANOVAs, t-tests and regression, require four assumptions to be made about the data they are applied to. These relate to the type of data, the independence of the data, the normality of the data, and the variance within the data. These four assumptions are described in Table 2. Additional assumptions may also be required for some types of tests. For example, linear regression has other assumptions such as no perfect linear relationship between two or more of the predictors ('multicollinearity'), and that the relationship between predictors and the predicted is linear. See Berry [1993] for further information about regression assumptions.

TABLE 2. Assumptions of parametric statistical tests.

Assumption	**Description**
Data are measured at least at interval level	The response or property being measured should be recorded using a dependent variable on an interval scale, minimum, or on a continuous scale. The intervals on the scale should represent differences of equal magnitude. For example, if a 1-5 rating scale is used to measure a participant's perceived brightness of a space, the difference in perceived brightness between ratings of 1 and 2 should be the same as it is between ratings of 4 and 5.
Data are independent	Data from one participant should not influence data from another participant, which can be addressed through randomisation in experimental design. In within-subjects designs, we do not expect the responses from the same participant to be independent, but responses between different participants in within-subjects designs should be independent. In regression analysis, the errors in the regression model should also be uncorrelated.
Data are normally distributed	The raw data within each condition approximates a normal distribution, or the residuals (individual minus the mean value) approximate a normal distribution, depending on the type of test being carried out.

Variance is the same throughout the data	When comparing more than one group of participants, each of these groups should have approximately equal variance. If carrying out a correlation, the variance of one of your variables should be stable at all levels of the other variable. This is known as homogeneity of variance, or homoscedasticity, particularly in relation to regression analysis. In within-subjects designs with three or more conditions, an assumption of sphericity is also made. Sphericity refers to the variances of the differences between pairs of conditions being equal across all combinations of conditions.

Assumptions about whether interval data are used and the independence of data should be assessed and confirmed during the experimental design phase of any research, for example through appropriate selection of measurement methods and randomisation of conditions. Assumptions about the normality and variance of data can only be assessed once data has been collected, and it is good practice to demonstrate that data meets these two assumptions before parametric statistical tests are used.

3.2 Assessment of normality

Confirming whether the data collected within a study sufficiently meets the assumption of a normal distribution should be seen as an informed judgement based on a series of diagnostic checks, rather than a definitive black and white decision. Note also that in regression analyses, it is the residuals (errors between the predicted and actual values) that are required to be normally distributed, not the actual variable values themselves. Normality of residuals may also be adequate for between-subjects ANOVAs and independent t-tests [Williams et al, 2013].

Three types of checks should be carried out to perform a comprehensive assessment of normality: 1) Visual inspection of graphical representations of the data; 2) Assessment of descriptive statistics; and 3) Statistical tests of deviation from a normal distribution. These methods are illustrated using two sets of simulated data, representing normal and non-normal distributions. The normally distributed data has been generated using the rnorm function within the R software package (version 3.4.0, R Core Team, 2017), with the parameters of sample size = 100, mean = 5, standard deviation = 1.5. The non-normal data are based on a positively-skewed exGaussian distribution. This type of distribution is frequently found in reaction time data [Palmer et al, 2011], and reaction times are commonly used as a response measure in lighting research [e.g. He et al, 1997; Fotios et al, 2017; Cengiz, Puolakka & Halonen, 2015]. The simulated non-normal data has been produced using the retimes package [Massidda, 2013], with the same parameters as for the normal data (sample size = 100, mean = 5, standard deviation = 1.5), and with the additional tau parameter, representing the exponential decay of the distribution tail, set at 4.

The distribution of a dataset can be visually inspected using three types of plot - a histogram, a quantile-quantile (Q-Q) plot, and a boxplot. The simulated normal and non-normal data have been plotted using these three types of visualisation in Figs. 1, 2 and 3.

The histogram represents a dataset by showing the counts of values within equally-sized ranges or ‘bins’. The size of these ranges, the ‘binwidth’, that is chosen for the histogram can have a large impact on the appearance of the data and its distribution. The binwidths chosen for the normal and non-normal data in Fig. 1 respectively were 0.81 and 1.47. These were selected using the Freedman-Diaconis rule of determining optimal bin size (see Equation 1). Other methods are also available for selecting the optimal binwidth, such as Sturges’ Rule and Bayesian optimal binning.

Equation 1 $(2 \times IQR) / n^{1/3}$

IQR = Interquartile range of data
n = sample size

<<< INSERT FIGURE 1 HERE >>>

Fig. 1. Histograms of simulated data with a normal (left) and non-normal (right) distribution.

Quantile-quantile plots compare actual data against data that would be expected if they were from a particular distribution (in this case, the normal distribution). Normally-distributed data would represent a straight line on the Q-Q plot and deviations away from this straight line indicate deviations away from a normal distribution. The nature of any divergence from a straight line can also reveal something about how the data fails to conform with normality. Figure 2 illustrates how the normal data follow a relatively straight line, whereas the non-normal data curve upwards at the larger response values, confirming the positive skew that is evident from the histogram.

<<< INSERT FIGURE 2 HERE >>>

Fig. 2. Quantile-quantile plots of simulated data with a normal (left) and non-normal (right) distribution.

A further method for visually inspecting the distribution of data is the box plot. The median value is indicated by the solid horizontal line within the box. The box itself represents values that are between the 25th and 75th quartiles. The vertical lines or whiskers show the extent of values that are within 1.5 times the IQR from each end of the box (greater than upper quartile + 1.5 IQR or less than lower quartile - 1.5 IQR). Values that are beyond this are shown as outliers and represented by individual data points. The boxplot would suggest a normal distribution if it was approximately symmetric overall, the median line was at the centre of the interquartile box, the whiskers are symmetric and slightly longer than the subsections of the interquartile box above and below the median line, and the number of outlying data points is small [Ghasemi & Zahediasl, 2012]. How small the number of outlying data points should be depends on the sample size. In a normal distribution, 0.8% of values would be expected to be more extreme than the upper or lower quartile ± 1.5 IQR and therefore flagged as an outlying value in the boxplot [Dawson, 2011]. Figure 3 shows boxplots for the simulated normal and non-normal datasets.

<<< INSERT FIGURE 3 HERE >>>
Fig. 3. Boxplot visualisations of the simulated normal (left) and non-normal (right) datasets.

Alongside visual inspections of the data, it is also useful to quantify any potential deviations from a normal distribution. A first approach to this is to quantify levels of skewness (how symmetrical the distribution is, and whether it has a number of extreme values that produce a long 'tail' to the distribution) and kurtosis (the relative thickness of the tails of the distribution, compared to a normal distribution). Statistical packages such as SPSS and R include simple methods for calculating skewness and kurtosis values. However the exact methods used in different packages may vary [Joanes & Gill, 1998]. Whatever method is used, a dataset with a pure normal distribution will have skewness and kurtosis values of zero.

To adequately assess deviations from normality it is necessary to convert the skewness or kurtosis statistic to a z-score by dividing it by its standard error. These transformed values can be compared against values you would expect to get by chance alone, based on a normal distribution. A z-score of ±1.96 is significant at $p < .05$, at ±2.58 it is significant at $p < .01$, and ±3.29 is significant at $p < .001$. A significant z-score may indicate the distribution has significant levels of skewness / kurtosis, although the threshold to use is a matter of judgement. As the sample size increases, the standard error becomes smaller, resulting in a larger z-score. Large samples are therefore more likely to provide transformed skewness and kurtosis statistics that appear significant, and it may therefore be appropriate to use a larger threshold to indicate whether the distribution of a large sample of data shows significant skewness or kurtosis. Field, Miles and Field [2012] suggest it is not appropriate to utilise z-score values of kurtosis and skewness for samples larger than 200. The z-score values of skewness and kurtosis for the simulated normal set of data are -0.30 and 0.23 respectively, indicating no evidence of skewness or kurtosis. The values for the non-normal dataset were 7.91 and 10.44, confirming significant skewness and kurtosis. These values have been calculated using the stats.desc function in the pastecs R package [Grosjean & Ibanez, 2014].

A further method for quantitatively assessing whether a distribution is normal or not is through use of a statistical test assessing a distribution for deviations from normality. The most commonly-used two tests are the Shapiro-Wilk test and the Kolmogorov-Smirnov test (other tests of normality are also available, such as the Anderson-Darling test, D'Agostino-Pearson omnibus test, and Jarque-Bera test). If the test produces a significant p-value, this indicates the data significantly deviates from a normal distribution. The Shapiro-Wilk may be a more sensitive and powerful test than other normality tests [Razali & Wah, 2011; Yap & Sim, 2011], although it is often assumed it is best used with samples less than 50 as the original development of the test by Shapiro and Wilk was limited to samples of this size or less [Shapiro & Wilk, 1965].

Statistical tests of deviation from normality suffer from over-sensitivity as the sample size increases, and may indicate even very minor deviations from normality as being significant with larger samples. This is illustrated in Fig. 4. This plots the probability that a Shapiro-Wilk test will give a significant result depending on the sample size, when the sample is taken from a population that shows marginal normality (ex-Gaussian distribution, n = 10,000, mean

= 5, standard deviation = 1.5, tau = 1). These probabilities were calculated using a Monte Carlo method in which 100 samples were drawn from the marginally-normal population for each sample size between 10 and 500. The calculated probability for each sample size was based on the proportion of Shapiro-Wilk tests that produced a significant ($p < .05$) result. As the Shapiro-Wilk test will produce a significant result even with a near-normal distribution, given a large enough sample, it may be appropriate to use a more stringent alpha with larger samples. When large samples are involved, consideration of the W statistic calculated by the Shapiro-Wilk test may also be useful in assessing whether any deviation from normality is problematic. Minor deviations from normality in a large sample will still produce a significant p-value, but if the W statistic is still large (e.g. greater than 0.98), the deviation could be considered to be minor and make little difference to the validity of a parametric statistical test.

<<< INSERT FIGURE 4 HERE >>>
Fig. 4. Probability that Shapiro-Wilk test will be significant for a sample taken from marginally-normal population, by sample size.

3.3 Assessment of equal variance

Parametric tests assume that the variance within different parts of your data are equal. If you are using a factorial design and have collected data across different groups, this means the variance within each of those groups should be approximately equal. If your data are not grouped but are continuous, for example in a design that uses regression, then the variance in data for one variable should be equal across all levels of the other variable [Field, Miles & Field, 2012].

When data are collected in different groups, for example recording ratings of spatial brightness for two or more different lamp types, the variance within these groups should be equal. Levene's test can be used to test whether this assumption of homogeneity of variance is true. A significant result on this test ($p < .05$) indicates that the variances of the different groups do significantly differ, and the assumption of equal variance is violated. However, as the sample size increases, Levene's test is more likely to flag even minor differences in variance as significant, and this may be inappropriate. Field, Miles & Field [2012] suggest also assessing homogeneity of variance using the Hartley's F_{max} statistic. This is the ratio of variance in the group with the largest variance to the group with the smallest variance. If this ratio is greater than a critical value for a given sample size, then the variances within the data are unlikely to be equal. The critical values are given in Pearson and Hartley [1954], but as a rule of thumb, a sample size of 10 per group would require an F_{max} of less than 10 to demonstrate equal variance, for a sample size of 15-20 per group F_{max} should be less than 5, and for sample sizes of 30-60 per group F_{max} should be less than 3 [Field, Miles & Field, 2012].

When a correlational design is used and data collected are continuous, for example as in linear regression, the assumption of equal variance (the variance on one variable is equal across all levels of another variable) should be checked using visual inspection methods, plotting the predicted value against its residual. Figure 5 gives an example of two such plots, one showing data that meets the assumption of equal variance the other showing data that

does not meet this assumption. Data from variables that have equal variance should present a random array of data points dispersed around zero. The data points form a funnel shape if the data comes from variables with unequal variance, indicating that the predictive power of the regression model systematically changes as the fitted value changes.

<<< INSERT FIGURE 5 HERE >>>
Fig. 5. Example plots of fitted values compared with residuals for linear regression model, showing data with equal variance (left) and unequal variance (right).

In within-subjects designs with more than three conditions, it is also important to check the assumption of sphericity. This assumes that the variances of the differences between pairs of conditions are the same across all possible pairs of conditions. Mauchly's test of sphericity is usually used to test this assumption. A significant result suggests a violation of the sphericity assumption.

3.4 Consequences of violating statistical test assumptions

To demonstrate the consequences of using parametric tests on data that does not meet test assumptions, a Monte Carlo procedure was used to identify the likelihood that a parametric test will produce a different conclusion to a non-parametric test when used on samples drawn from populations that are not normally distributed.

Two populations of simulated reaction times measured in milliseconds with ex-Gaussian distributions were generated using the pastecs R package. Both populations had N = 1000, and parameters of standard deviation = 250 and tau = 500. Population 1 were given a mean parameter of 500, and population 2 a mean parameter of 850. The distributions of populations 1 and 2 are shown in Fig. 6, illustrating a clear difference in reaction times between the two populations. Two random samples of n = 15, one from each of these populations, were drawn and tested for normality using the Shapiro-Wilk test. As the aim was to ensure one of these paired samples were not normally distributed, if neither of the samples produced a significant result ($p < .05$) on the normality test, the samples were discarded and new samples drawn from each of the populations. This process was stopped when a thousand pairs of samples from each of the populations were obtained. The parametric independent t-test and the non-parametric Mann-Whitney U-test were used to compare the samples in each of these pairs. These tests provided discrepant conclusions for 17% of the sample pairs, defined as disagreement about whether the samples were significantly different at $p < .05$. The large majority of these disagreements (82%) occurred because the Mann-Whitney test indicated a significant difference whereas the t-test did not. This simple demonstration highlights the potential that use of parametric methods will provide inappropriate conclusions when used on non-normal data, in comparison to conclusions drawn from appropriately used non-parametric methods. This illustration highlights the potential for increased risk of incorrectly accepting the null hypothesis (Type II error) when using parametric methods on non-normal data. However, in different circumstances there is also potential for an increased risk of making a Type I error [Wilcox, 1998].

<<< INSERT FIGURE 6 HERE >>>
Fig. 6. Density plots showing distributions of two populations of simulated reaction time data, generated using the pastecs package in R.

3.5 Addressing violations of assumptions

If data are assessed as violating one or more of the assumptions required by parametric tests, there are three options:

1) Accept the violation and proceed with using a parametric test anyway. The magnitude of the violation should be considered. If small, it may be acceptable to use a parametric test such as an ANOVA. For example, Field, Miles and Field [2012] suggest the ANOVA is reasonably robust to violations of homogeneity of variance when group sizes are equal, and minor violations of the normal distribution can also still produce results that are similar to those when the data are normally distributed [Schmider et al, 2010]. The assumption of normality can be ignored with increasing confidence as the sample size rises above 30. This is due to the Central Limit Theorem, which states that the means of a random set of samples from a population (the sampling distribution) approaches normality as the size of the sample increases. Sample sizes of $N \geq 30$ generally produce a normal distribution of sample means for all but the most extreme non-normal distributions. For such distributions, a sample size greater than 30 may be required for the sampling distribution to approach normality, but such extreme distributions are rare in most research. If the sampling distribution is normal, the assumption of normality is met and we can proceed with using a parametric test even if the distribution of our sample is not normal. However, with group samples of $N < 30$ that do not follow a normal distribution, or if other parametric test assumptions are violated, the implications of proceeding with parametric testing, as highlighted in section 4.4, should be carefully considered.
2) Transform the data to make it meet the violated assumption. Most dependent variables are measured on a linear scale, but the use of a linear scale is arbitrary and data can be legitimately transformed using some function to address the violation of a parametric assumption whilst maintaining the informational integrity of the data. Transforms can be used to address violations of different assumptions, for example to produce a normally-distributed set of data, or to increase homogeneity of variance. It is important to be aware that although transforming data does not change the relationship between different variables, it does change the differences between variables. This means when comparing differences within the same variable (e.g. responses on each level of a within-subjects factor), you should transform all levels of that variable, not just those that violate one of the assumptions you are assessing [Field, Miles & Field, 2012]. The type of data transformation used will depend on how an assumption is violated. For example, log, square root or reciprocal transformations can correct for positive skew and unequal variances, whilst a reverse score transformation can correct for a negative skew. Further details about transformations can be found in Field, Miles & Field (2012) and McDonald [2014].
3) Use an alternative non-parametric statistical test that is robust to the violation. A range of statistical tests exist that do not rely on parametric assumptions. A summary of some non-parametric alternatives to commonly used parametric tests is given in Table 3. Before deciding to use a non-parametric test, it is worth noting that they generally have

less power than parametric tests, leading to an increased risk of making a Type II error (falsely retaining the null hypothesis). Other methods that are robust to violations of parametric assumptions but not listed in Table 3 are also available. For example, if variance within groups is not equal, Welch's separate variances t-test (for comparing two groups of data) or Welch's F test (for comparing three or more groups of data) are available. Some researchers have even suggested abandoning the traditional Student's t-test in favour of Welch's t-test, as it performs better when sample sizes and variances are unequal between groups, or just as well when they are equal [Delacre, Lakens & Leys, 2017]. Bootstrapping procedures, and generalised linear mixed models also provide alternative approaches if data fail to meet parametric assumptions. Some tests also provide corrected statistics to account for violated assumptions. For example, if sphericity is not present in the data, a correction can be applied to produce a valid F-ratio. Options include the Greenhouse-Geisser correction and the Huynh-Feldt correction. Further details about applying these corrections can be found in most statistics text books, e.g. Field, Miles & Field (2012). Further information about non-parametric statistics is available in Conover [1999] and Siegel and Castellan [1988].

TABLE 3. Non-parametric alternatives to commonly used parametric tests (see Motulsky, 1995).

Assessment being made	**Parametric test**	**Non-parametric test**
Compare one group to a hypothetical value	One-sample t-test	Wilcoxon signed-rank test
Compare two independent groups	Independent t-test	Mann-Whitney test
Compare three or more independent groups	One-way ANOVA	Kruskal-Wallis test
Compare two dependent groups (within-subjects)	Dependent t-test	Wilcoxon signed-rank test
Compare three or more dependent groups	Repeated-measures ANOVA	Friedman test
Association between two variables	Pearson correlation	Spearman correlation
Predict value based on another value	Linear / Nonlinear regression	Nonparametric regression / logistic regression

4. Reporting of effect sizes

When conducting research, we are generally interested in discovering whether our variables of interest have some effect on what we are studying. This effect may relate to a difference between groups, for example hazard detection rates under different lighting conditions. Alternatively, it may relate to associations between variables, for example whether outdoor illuminance levels are associated with perceived safety. If applied appropriately (see Section 3), Null Hypothesis Statistical Testing and the p-value produced can provide evidence towards an effect being present (or at least that no effect, the null hypothesis, is implausible). As well as knowing whether an effect may be present, we are also interested in how big this

effect is – do our variables have a big influence on what we are measuring, or only a trivial influence? A range of methods are available to calculate the size of an effect, some of these are listed in Table 4. Measures of effect size often produce a standardised value which allows comparison between studies using different metrics and a consistent 'language' of effect magnitudes. Further information about effect sizes and their calculation is available elsewhere (e.g. Lakens, 2013; Cohen, 1988, 1992; Sullivan & Feinn, 2012).

The size of any effect revealed within a study is a valuable piece of information when results are reported, for three reasons [Lakens, 2013]. First it provides information about the magnitude of the effect found, allowing its practical importance to be considered. This information cannot be adequately gleaned from only a p-value [Durlak, 2009]. Second, it can be incorporated into meta-analyses that combine the findings from multiple studies to provide holistic evidence and more definitive conclusions about a research question or area. Third, it can be used in the design of future related research to estimate required samples sizes, through a priori power analyses, as discussed above. However, despite the evidential and scientific value of reporting effect sizes, this is rarely done in lighting research. The review of recent lighting research papers, and papers related to spatial brightness (Section 2), showed that only 24% of the 50 studies included in the review reported effect sizes of some kind, with the majority of effect size measures being R^2 values from a linear regression.

It may be possible for the reader of a study article to calculate for themselves some measures of effect size using commonly reported data such as the means and standard deviations (Cohen's d can be estimated using the difference between group means and the pooled standard deviation, for example). However most readers are unlikely to make such calculations for every study they read about, or may not have sufficient information available to make the calculations correctly. It is also unwise to rely on the reader's intuitions about statistical effect sizes or power achieved by a study, even if they are highly statistically literate, as they are likely to be incorrect [Tversky & Kahneman, 1971; Bakker et al, 2016]. Authors of lighting research papers should therefore be encouraged to explicitly report effect sizes within their results. The type of effect size measure that should be reported will depend on the statistical test used and the experimental design (e.g. see Table 4 for effect size measures associated with statistical tests reported in lighting research papers included in the review). The range of possible effect size measures limits any discussion of how to calculate and report effect sizes in this paper, but a number of relevant guides exist (e.g. see Lakens, 2013, and Durlak, 2009). G*Power [Erdfelder et al, 1996] is also recommended as a convenient and powerful open-source application that can calculate effect sizes for a range of tests and designs.

The limited reporting of effect sizes within lighting research literature may reflect wider inconsistencies in how statistical information is reported, as indicated by the review carried out in Section 2. Variations were found in the reporting of summary statistics, including whether measures of variance such as standard deviations or standard errors were presented. There were also inconsistencies in the reporting of test statistics - some papers provided the test statistic, such as t or F, the degrees of freedom, and the p-value, whilst others provided only the p-value. In many circumstances, p-values were not even provided, particularly when a test was not-significant. There were also variations in the precision of statistical reporting and whether exact p-values were given, with the broad statement of 'p <

.05' being frequently used. Whether a statistical test was one- or two-tailed was rarely stated.

5. Sample size and power

As discussed in previous sections, key goals of any research study is to discover whether an effect exists (which requires appropriate application of statistical tests, see Section 3) and the magnitude of any effect (which requires the calculation and reporting of a measure of effect size, see Section 4). The sample size used has implications for both these objectives. Sample size is an essential determinant of the size of the effect that study will be able to reveal. It also contributes to determining the power of the study - the probability that a significant effect will be revealed through statistical testing when a true effect does really exist (i.e. the probability of avoiding a Type II error). Increasing the sample size increases the power of a study thus making it more able to detect an effect of a smaller size, reducing the likelihood that the null hypothesis will be incorrectly accepted. Figure 7 shows how power changes with sample size, for small, medium and large effects, using independent and dependent t-tests. Note that the effect size metric for between-subjects (left plot) is Cohen's d, whilst for within-subjects (right plot) it is Cohen's d_z.

<<< INSERT FIGURE 7 HERE >>>

Fig. 7. Calculated power of one-tailed independent t-test for between-subjects designs (left) and dependent t-test for within-subjects designs (right), by sample size and effect size (Cohen's d for between-subjects, Cohen's d_z for within-subjects), assuming an alpha of 0.05. Independent t-test uses group sample size, not total sample size.

Within-subjects designs have greater power than between-subjects designs for the same sample sizes due to reduced individual variance. A number of effect size measures exist for within-subjects designs (e.g. see Lakens, 2013, and Rosnow & Rosenthal, 2003). For example, the classical Cohen's d can be used for two matched groups, however this does not take account of the correlations between paired values in within-subjects data. An alternative is to calculate Cohen's d_z which accounts for the correlated nature of paired data. The greater the correlation between the paired values, the larger d_z becomes. Caution should be taken in comparing the size of d (for between-subjects designs) and d_z (for within-subjects designs) however. For a given difference between two means and associated standard deviations in values, the effect size calculated if the data are treated as within-subjects (d_z) is likely to be considerably larger than if the data are treated as between-subjects (d). The size of this difference will depend on the strength of the correlations between paired values. Lakens (2013, p. 7-8) provided hypothetical analysis to illustrate this, showing that when illustrative data were treated as between-subjects, Cohen's d was 1.13, but when the same data were treated as within-subjects, Cohen's d_z was 1.50. Effect sizes for within-subjects designs may therefore be "inflated", relative to the default effect size calculations for between-subjects designs [Dunlap et al, 1996].

If comparing two unrelated groups using an independent t-test, a sample size of 310 participants in each group would be required to find an effect size of d = 0.2 (a small effect according to Cohen's thresholds), assuming a power of 0.8. If comparing two related groups

using a dependent t-test, a sample size of 156 would be required to find an effect size of d_z = 0.2. The inflated nature of the d_z measure means it may be more realistic to seek a larger effect size. For illustration purposes, a d_z of 0.4 would require a sample size of 41. The median sample sizes found in the review of lighting publications (Section 2) were all below 40, highlighting the potential that ongoing research in the lighting field runs the risk of being underpowered, if past sample sizes are indicative of sample sizes used in future research.

To explore this issue in more detail estimates of the power capable of being achieved by the sample sizes and statistical tests used in papers included in the review (section 2) were calculated for small, medium and large effect sizes. Note that this is not an attempt to calculate the observed power within each study. Post hoc calculation of observed power, using the observed effect size and sample size used, provides almost no information of value. By definition, a study had sufficient power to detect an effect if a significant effect was revealed. As Hoenig and Heisey stated: “Power calculations tell us how well we might be able to characterize nature in the future given a particular state and statistical study design, but they cannot use information in the data to tell us about the likely states of nature” [Hoenig & Heisey, 2001, p. 23]. In this analysis, the sample sizes and statistical tests reported in a sample of lighting research papers are used as example data in determining the power achieved for different effect sizes. This aims to reveal the power capable of being achieved by existing research practices within the lighting field.

The effect size criteria used in this analysis were defined by convention for the specific statistical test [e.g. Cohen, 1988, 1992; Olivier, May & Bell, 2017] as shown in Table 4. Power estimates were calculated for each type of test and sample size used within these papers, and for each threshold of effect size (small, medium and large), using the G*Power software [Erdfelder et al, 1996]. If the same type of test was used multiple times in the same paper, only the test parameters and sample size that would produce the largest power were included in this analysis. This meant each paper provided only one set of details per category of test carried out, providing a more representative sample of values and avoiding some studies that used large numbers of tests dominating the results of this analysis.

For the six papers that did not report inferential statistics, a judgement was made about an appropriate statistical test and power estimated based on this. A total of 67 estimates of power for each effect size criterion were calculated, from 42 papers. Power estimates were unable to be estimated from eight papers due to the type of test used (e.g. Friedman’s ANOVA) or insufficient information provided. The distribution of power estimates are shown in Fig. 8. These histograms illustrate how current lighting research practice, in terms of the sample sizes and statistical tests used, is very unlikely to be capable of revealing a small effect, with no statistical tests reaching the recommended power criteria of 0.8. For detection of a medium-sized effect, only 42% of studies reported tests that would reach the 0.8 power criteria. The situation was better for detection of a large effect, with 75% of reported tests capable of reaching a power of 0.8 or more.
Note also that the power estimates suggested in Fig. 7 and Fig. 8 may be optimistic as they are based on use of a one-tailed test. One-tailed tests, in which the direction of an effect is explicitly predicted, provide greater power than two-tailed tests. The difference in power is a function of a number of factors including the test used, the effect size and the sample size. As an example, the power for a group sample size of 20 and an effect size of 0.2 is 0.15 for a one-tailed dependent t-test. This reduces to 0.09 when a two-tailed t-test is used. Lighting

research papers included in the review rarely reported whether a one- or two-tailed test was used. Justification for use of a one-tailed test was also absent in nearly all reviewed papers, a finding that is in common with other research areas [e.g. Ruxton & Neuhauser, 2010].

Although the ability to detect small effect sizes in lighting research appears to be limited, the need or desirability of detecting small effects should also be considered. It is reasonable, particularly in applied lighting research, for investigators to only be interested in detecting effect sizes of a certain magnitude. The practical implications of a small effect may be negligible, and this could justify powering studies to only detect larger effect sizes.

TABLE 4. Small, medium and large effect size criteria conventions, for different effect size measures, based on statistical tests used in studies published in Lighting Research & Technology and LEUKOS in 2017.

Effect size statistic	**Statistical test/s**	**Small effect**	**Medium effect**	**Large effect**
Cohen's d	One-sample t-test	0.2	0.5	0.8
Cohen's d_z	Dependent t-test; Wilcoxon signed-rank test	0.2	0.5	0.8
Cohen's f	Repeated-measures ANOVA; Between-subjects ANOVA	0.1	0.25	0.40
Cohen's f^2	Regression	0.02	0.15	0.35
Odds ratio	McNemar test	1.22	1.86	3.00
Kendall's w	Friedman ANOVA	0.1	0.3	0.5
g	Binomial test	0.05	0.15	0.25

It is good practice to carry out an a priori power analysis to determine the sample size required to be confident in revealing an effect if there is one truly present. Despite its benefits, no evidence was found of a priori power analysis in the sample of lighting research papers reviewed, suggesting it may not be routine practice in lighting research.

A power analysis requires knowledge of three things. The first is the alpha level, the probability of observing the measured effect you are willing to accept, when in reality no true effect exists (effectively, the probability of making a Type I error). Common practice usually sets the alpha at 0.05, although as highlighted earlier, the choice of alpha to use should not be inflexible [Lakens et al, 2017]. The second thing we need to know is the power we aim to achieve with our test - the probability of detecting an effect when one truly does exist (avoiding making a Type II error). A common minimum required power is 0.8 [Cohen, 1992]. The final piece of information required for a power analysis is the effect size that is anticipated, or that the test should be capable of revealing. Armed with these three pieces of

information we can calculate required sample sizes, using statistical software such as G*Power [Erdfelder et al, 1996] or the pwr package in R.

The alpha and power threshold are generally predetermined based on conventions, but a potential effect size has to be estimated. One approach to determining an estimated effect size is by examining previous related literature to estimate an average effect size for the type of effect you are interested in. This may be difficult within lighting research as there are very few meta-analyses that summarise effect sizes from a range of studies within a specific research topic, and many published studies fail to report effect sizes (as demonstrated in the review carried out for this paper) or provide the necessary statistics to calculate an effect size. There is also the potential that the effect sizes reported in published literature may not reflect the true effect size due to publication bias and the general under-powering of studies [e.g. Button et al, 2013; Paterson et al, 2013; Quintana, 2017]. An alternative approach to deciding on an effect size for use in a power analysis is to state the minimum effect size you are willing to accept as detectable with your study, or assess what the minimum effect size would be for it to be meaningful and not trivial (the SESOI – Smallest Effect Size Of Interest. Albers & Lakens, 2018).

<<< INSERT FIGURE 8 HERE >>>

Fig. 8. Estimated power of statistical tests used in reviewed papers, based on sample size and other parameters such as number of measurements (in within-subjects methods), for small (top), medium (centre) and large (bottom) effect sizes. Vertical dashed line indicates conventional minimum recommended power of 0.8 [Cohen, 1988].

6. Conclusions

Publication bias and the reproducibility crisis are issues that pose a significant risk to the evidential value of research within a number of fields, but particularly within lighting research. At the heart of these issues lies the risk of making Type I or Type II errors. The statistical methods employed in research are designed to reduce these errors, and their role in determining the presence and importance of any effect are critical to the veracity of published research. This paper reviewed a sample of general and topic-specific lighting research papers. The review highlighted the relatively small samples used in behavioural lighting research, and the lack of power this introduces. The sample sizes used in most lighting studies may only be capable of revealing medium to large effects.

It is important to consider whether an effect of a certain size is of practical significance. Depending on the specific research area and question being investigated, a small effect size may be insufficiently interesting or noteworthy to warrant investigation, and researchers may only be interested in discovering effects equal to or greater than a certain magnitude. With limited research funding and resources available, the size of an effect that is worth detecting is an important consideration when determining the sample size of a study. Whatever size of effect is judged to be sufficiently large to be of interest, it remains important to justify the sample size used. However, the justification of sample sizes, based on anticipated or targeted effect sizes, was virtually non-existent within the papers reviewed here.

One possible reason for this absence of sample size justification is that the practice of reporting effect sizes in lighting research papers is not commonplace, and therefore it may be difficult to estimate anticipated effect sizes with any confidence. Only 24% of reviewed papers reported any kind of effect size measure. The American Psychological Association Task Force on Statistical Inference [Wilkinson, 1999] states that: *"...reporting and* interpreting effect sizes in the context of previously reported effects is essential to good research." (p.599). Increased reporting of effect sizes should be encouraged within lighting research, as should detailed, accurate and appropriate statistical analysis and reporting. This can help reduce the promotion of unsupported findings within lighting research literature.

The review presented in Section 2 highlighted that parametric statistical tests, including t-tests, ANOVAs and linear regressions, are the dominant type of testing carried out. Parametric tests require a number of assumptions to be made about the data, including a normal distribution and equality of variances. Despite this very few papers explicitly stated these assumptions had been assessed before a statistical test was used. Inappropriate use of parametric statistical tests can result in an increase in Type II errors (false negatives), as illustrated by the simple example presented in Section 3.4. Wilcox [1998] also demonstrated that even a small departure from normality could reduce the power of a t-test from .96 to .28. Inappropriate use of parametric tests may also lead to an increase in Type I errors, as stated by Erceg-Hurn & Mirosevich [2008]: *"... the p values reported by statistical packages such as* SPSS may be extremely inaccurate if the data being analyzed are non-normal and/or heteroscedastic; the inaccuracy may lead researchers to unwittingly make Type I *errors"* (p. 593). The evidential value and accuracy of studies within the lighting research literature would be improved if the assumptions of the statistical tests proposed for use were assessed and reported on. This paper provides guidance on how the assumptions of a normal distribution and equal variances can be assessed.

This paper highlights three issues relevant to improving the evidential quality within lighting research - determination and justification of sample sizes, assessment of test assumptions, and reporting of statistical results particularly effect sizes. Further treatment of these issues can also be found in a number of other texts [e.g. Cohen, 2013; Haslam & McGarty, 2018; Abelson, 1995; Field, Mils & Field, 2012]

There are other practices and methods that can improve evidential quality that also warrant discussion within the lighting research community. For example, preregistration of studies may help address publication bias, control for researcher degrees of freedom [Simmons,

Nelson & Simonsohn, 2011] and ‘Questionable Research Practices’ [John, Loewenstein & Prelec, 2012]. Research quality and transparency can also be improved through justification of all research design decisions within a study, including the sample size used, analytical methods, and p-value thresholds chosen [Lakens et al, 2017]. Some researchers suggest abandoning the term ‘statistically significant’ [Lakens et al, 2017; McShane et al, 2017] as it induces a rigid interpretation of a set of results when in reality the interpretation may be context-dependent and the meaning of ‘significance’ may vary depending on the topic and research field. Discussion of these ideas would be valuable in the context of lighting research. However, null hypothesis statistical testing is likely to remain the de rigeur method for assessing results in the foreseeable future. All those involved in research publication, from researchers to reviewers and editors, should aim to ensure this approach is applied appropriately, taking heed of the three issues discussed in this paper. This should include accurate and appropriate reporting of the results of such statistical analysis. To support this aim lighting journals should consider adopting existing guidelines for reporting quantitative results provided by expert organisations such as the American Psychological Association [Appelbaum et al, 2018], for studies with outcome measures from the behavioural and social sciences.

Appropriate statistical analysis and reporting will help ensure research resources are not wasted, participants’ time is not wasted and they are not exposed to undue risk through participation in unnecessary or poor research, and readers’ time is not wasted.

Funding
This work was carried out with support from the Engineering and Physical Sciences Research Council (EPSRC) grant number EP/M02900X/1.

Disclosure Statement
The author reported no declarations of interest.

References

Abelson, R. P. 1995. Statistics as Principled Argument. Hillsdale, NJ: Lawrence Erlbaum Associates.

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. Journal of Experimental Social Psychology, 74, 187-195.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. American Psychologist, 73(1), 3-25.

Bakker, M., & Wicherts, J. M. 2011. The (mis) reporting of statistical results in psychology journals. Behavior Research Methods, 43(3), 666-678.

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. 2016. Researchers' intuitions about power in psychological research. Psychological Science, 27(8), 1069-1077.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. 2017. Redefine statistical significance. Nature Human Behaviour, 1.

Berry, W. D. (1993). Understanding regression assumptions. Sage university paper series on quantitative applications in the social sciences, 07-092. Newbury Park, CA: Sage.

Boyce, P. R., & Cuttle, C. 1990. Effect of correlated colour temperature on the perception of interiors and colour discrimination performance. Lighting research & technology, 22(1), 19-36.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365-376.

Cengiz, C., Puolakka, M., & Halonen, L. 2015. Reaction time measurements under mesopic light levels: Towards estimation of the visual adaptation field. Lighting Research & Technology, 47(7), 828-844.

Cohen, B. H. 2013. Explaining Psychological Statistics (4th ed.). New York: Wiley.

Cohen, J. 1988. Statistical power analysis for the beahavioral sciences. New York: Lawrence Erlbaum Associates.

Cohen, J. 1992. A Power Primer. Psychological bulletin, 112(1), 155-159.

Conover, W. J. 1999. Nonparametric Statistics, 3rd edition. New York: John Wiley and Sons.

Davis, R. G., & Ginthner, D. N. 1990. Correlated color temperature, illuminance level, and the Kruithof curve. Journal of the Illuminating Engineering Society, 19(1), 27-38.

Dawson, R. 2011. How significant is a boxplot outlier? Journal of Statistics Education, 19(2). Available online: http://ww2.amstat.org/publications/jse/v19n2/dawson.pdf [Accessed 14/02/2018].

Delacre, M., Lakens, D., & Leys, C. 2017. Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. International Review of Social Psychology, 30(1), 92-101.

DiLaura, D. L., Houser, K. W., Mistrick, R. G., & Steffy, G. R. 2011. The Lighting Handbook (10th ed.). New York: IESNA.

Dorey, F. 2010. In brief: The P value: What is it and what does it tell you? Clinical Orthopaedics and Related Research, 468(8), 2297-2298.

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. 2017. Low statistical power in biomedical science: a review of three human research domains. Royal Society Open Science, 4(2), 160254.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods, 1(2), 170-177.

Durlak, J. A. 2009. How to select, calculate, and interpret effect sizes. Journal of Pediatric Psychology, 34(9), 917-928.

Erceg-Hurn, D. M., & Mirosevich, V. M. 2008. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. American Psychologist, 63(7), 591-601.

Erdfelder, E., Faul, F., & Buchner, A. 1996. GPOWER: A general power analysis program. Behavior Research Methods, Instruments, & Computers, 28, 1–11.

Fanelli, D. 2010. "Positive" results increase down the hierarchy of the sciences. PloS one, 5(4), e10068.

Field, A, Miles, J, Field, Z. 2012. Discovering Statistics Using R. London: Sage Publications.

Fisher, R. A. 1925. Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd.

Flynn, J. E., Hendrick, C., Spencer, T., & Martyniuk, O. 1979. A guide to methodology procedures for measuring subjective impressions in lighting. Journal of the Illuminating Engineering Society, 8(2), 95-110.

Fotios, S., & Goodman, T. 2012. Proposed UK guidance for lighting in residential roads. Lighting Research & Technology, 44(1), 69-83.

Fotios, S., Atli, D., Cheal, C., Houser, K., & Logadóttir, Á. 2015. Lamp spectrum and spatial brightness at photopic levels: A basis for developing a metric. Lighting Research & Technology, 47(1), 80-102.

Fotios, S. 2017. A revised Kruithof graph based on empirical data. LEUKOS, 13(1), 3-17.

Fotios, S., Cheal, C., Fox, S. & Uttley, J. 2017. The effect of fog on detection of driving hazards after dark. Lighting Research & Technology. Advance online publication, doi: 10.1177/1477153517725774.

García-Berthou, E., & Alcaraz, C. 2004. Incongruence between test statistics and P values in medical papers. BMC medical research methodology, 4(1), 13.

Ghasemi, A., & Zahediasl, S. 2012. Normality tests for statistical analysis: a guide for non-statisticians. International Journal of Endocrinology and Metabolism, 10(2), 486-489.

Glass, G. V. 1976. Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Grosjean, P., & Ibanez, F. 2014. Pastecs: Package for Analysis of Space-Time Ecological Series. R package version 1.3-18. https://CRAN.R-project.org/package=pastecs.

Haslam, A., & McGarty, C. (2018). Research Methods and Statistics in Psychology (3rd ed.). London: Sage.

He, Y., Rea, M., Bierman, A., & Bullough, J. 1997. Evaluating light source efficacy under mesopic conditions using reaction times. Journal of the Illuminating Engineering Society, 26(1), 125-138.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. The American Statistician, 55(1), 19-24.

Hubbard, R., & Ryan, P. A. 2000. The historical growth of statistical significance testing in psychology—And its future prospects. Educational and Psychological Measurement, 60(5), 661-681.

Ioannidis, J. P. 2005. Why most published research findings are false. PLoS medicine, 2(8), e124.

Jefferson, T., Alderson, P., Wager, E., & Davidoff, F. 2002. Effects of editorial peer review: a systematic review. Jama, 287(21), 2784-2786.

Joanes, D. N., & Gill, C. A. 1998. Comparing measures of sample skewness and kurtosis. Journal of the Royal Statistical Society: Series D (The Statistician), 47(1), 183-189.

John, L. K., Loewenstein, G., & Prelec, D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science, 23(5), 524-532.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. 2009. Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. Personality and Social Psychology Bulletin, 35(9), 1131-1142.

Kruithof, A. A. 1941. Tubular luminescence lamps for general illumination. Phillips Technical Review, 6, 65-73.

Lakens, D. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 863.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... & Buchanan, E. M. (2018). Justify your alpha. Nature Human Behaviour, 2(3), 168-171.

Lemoine, N. P., Hoffman, A., Felton, A. J., Baur, L., Chaves, F., Gray, J., ... & Smith, M. D. 2016. Underappreciated problems of low replication in ecological field studies. Ecology, 97(10), 2554-2561.

Massidda, D. 2013. retimes: Reaction Time Analysis. R package version 0.1.2. https://CRAN.R-project.org/package=retimes.

McDonald, J. H. 2014. Handbook of Biological Statistics (3rd ed.). Baltimore, Maryland: Sparky House Publishing.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. 2017. Abandon statistical significance. arXiv preprint arXiv:1709.07588.

Motulsky, H. 1995. Intuitive Biostatistics. New York: Oxford University Press.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. 2016. The prevalence of statistical reporting errors in psychology (1985–2013). Behavior Research Methods, 48(4), 1205-1226.

Olivier, J., May, W. L., & Bell, M. L. 2017. Relative effect sizes for measures of risk. Communications in Statistics-Theory and Methods, 46(14), 6774-6781.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science, 3496251, aac4716.

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. 2011. What are the shapes of response time distributions in visual search?. Journal of Experimental Psychology: Human Perception and Performance, 37(1), 58.

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. 2016. An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. Journal of Leadership & Organizational Studies, 23(1), 66-81.

Pearson, E. S., & Hartley, H. O. 1976. Biometrika tables for statisticians, Volume I (3rd ed.). New York: Cambridge University Press.

Quintana, D. S. 2017. Statistical considerations for reporting and planning heart rate variability case-control studies. Psychophysiology, 54(3), 344-349.

Razali, N. M., & Wah, Y. B. 2011. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21-33.

Rea, M.S. & Illuminating Engineering Society of North America 1993. IESNA Lighting Handbook, 8th ed. New York: IESNA.

R Core Team 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rosnow, R.L. & Rosenthal, R.R. 2003. Effect sizes for experimenting psychologists. Canadian Journal of Experimental Psychology, 57(3), 221-237.

Rothman, K. J. 2014. Six persistent research misconceptions. Journal of general internal medicine, 29(7), 1060-1064.

Ruxton, G. D., & Neuhäuser, M. 2010. When should we use one-tailed hypothesis testing?. Methods in Ecology and Evolution, 1(2), 114-117.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. 2010. Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. Methodology, 6, 147-151.

Shapiro, S.S., & Wilk, M.B. 1965. An analysis of variance test for normality (complete samples). Biometrika, 52, 591-611.

Siegel, S.C., & Castellan, N. J. 1988. Nonparametric statistics for the behavioural sciences. New York: McGraw-Hill.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11), 1359-1366.

Simons, R. H., Hargroves, R. A., Pollard, N. E., & Simpson, M. D. 1987. Lighting criteria for residential roads and areas. CIE, Venice, 274-277.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. 1995. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. The American Statistician, 49(1), 108-112.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. Journal of Graduate Medical Education, 4(3), 279-282.

Thiese, M. S., Arnold, Z. C., & Walker, S. D. 2015. The misuse and abuse of statistics in biomedical research. Biochemia Medica, 25(1), 5-11.

Tversky, A., & Kahneman, D. 1971. Belief in the law of small numbers. Psychological Bulletin, 76(2), 105-110.

Veitch, J. A. 2001. Psychological processes influencing lighting quality. Journal of the Illuminating Engineering Society, 30(1), 124-140.

Ware M. 2008. Peer Review: Benefits, Perceptions and Alternatives. PRC Summary Papers, 4:4-20.

Wilkinson, L. 1999. Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54(8), 594-604.

Wilcox, R. R. 1998. How many discoveries have been lost by ignoring modern statistical methods?. American Psychologist, 53(3), 300-314.

Williams, M. N., Gómez Grajales, C. A., & Kurkiewicz, D. (2013). Assumptions of Multiple Regression: Correcting Two Misconceptions. Practical Assessment, Research & Evaluation, 18 (11).

Yap, B. W., & Sim, C. H. 2011. Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation, 81(12), 2141-2155.