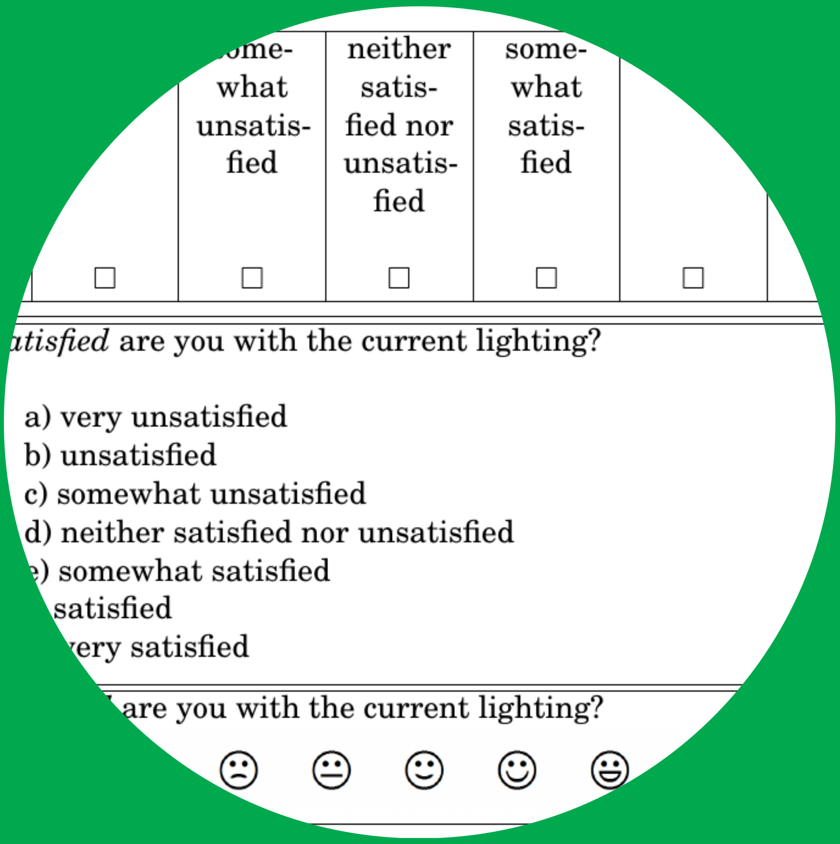
Mikko Hyvärinen Methodological questions in lighting acceptance and preference studies Aalto University

Department of Electrical Engineering and Automation

Methodological questions in lighting acceptance and preference studies

Mikko Hyvärinen



	omewhat unsatisfied	neither satisfied nor unsatisfied	somewhat satisfied	
□	□	□	□	□

atisfied are you with the current lighting?

a) very unsatisfied
b) unsatisfied
c) somewhat unsatisfied
d) neither satisfied nor unsatisfied
) somewhat satisfied
satisfied
ery satisfied

are you with the current lighting?

☹ 😐 ☺ ☺ 😃

Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 15/2015

Methodological questions in lighting acceptance and preference studies

Mikko Hyvärinen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall S1 of the school on 13 February 2015 at 12 noon

Aalto University
School of Electrical Engineering
Department of Electrical Engineering and Automation
Lighting Unit

Supervising professor
Prof. Liisa Halonen

Thesis advisor
Dr. Marjukka Puolakka

Preliminary examiners
Prof. Bert Weijiters, Ghent University, Belgium
Dr. Teija Vainio, University of Tampere, Finland

Opponent
Prof. Markku Turunen, University of Tampere, Finland

Aalto University publication series
DOCTORAL DISSERTATIONS 15/2015

ISBN 978-952-60-6065-1 (printed)
ISBN 978-952-60-6066-8 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6066-8

Unigrafia Oy
Helsinki 2015

Finland

NORDIC ECOLABEL
441 697
Printed matter

A? Aalto University

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Abstract

Author
Mikko Hyvärinen

Name of the doctoral dissertation
Methodological questions in lighting acceptance and preference studies

Publisher School of Electrical Engineering

Unit Department of Electrical Engineering and Automation

Series Aalto University publication series DOCTORAL DISSERTATIONS 15/2015

Field of research Illumination engineering

Manuscript submitted 12 September 2014 **Date of the defence** 13 February 2015

Permission to publish granted (date) 19 December 2014 **Language** English

☒ **Monograph** □ **Article dissertation (summary + original articles)**

Abstract

Investigating subjective aspects of lighting, such as preference and acceptance, involves evaluations and ratings by human subjects. The data from the subjects is gathered with questionnaire techniques.

This work reviews the literature published on lighting field on methodology for subjective lighting studies. The review continues to questionnaire research in general, and discusses several factors which may influence the results or the quality of the data, such as the form of response options, labelling the response alternatives, the number of response option categories, whether or not there should be an option for a neutral answer.

Response styles are addressed by a literature review, and a discussion on how they may affect a lighting acceptance or preference study.

The work reviews also a number persisting controversies related to statistical analysis of questionnaire data: null hypothesis significance testing, adjusting the significance criterion for multiple comparisons, and analysing ordinal data with parametric statistical methods. A review of arguments from both sides for each controversy is presented, and the reasons why these controversies persist are discussed

Twenty subjective lighting studies published since 2005 are analysed for the methodology used and the way the methodology and results are presented. The finding was that about half of the studies considered did not report their methods, analysis or results in sufficient detail. The statistical power is also often quite low.

Two experiments were conducted to see how varying the response format affects the results gained from a lighting experiment, and to evaluate if the choice of parametric or non-parametric statistical method effects the conclusion. The findings were, that reversing the response option order did not affect the conclusion or characteristics of the data, but providing labels for every response option resulted into a slightly higher scores than when only the end-points were labelled. Both parametric and non-parametric statistical methods led to the same conclusion, and thus this work does not support the view that parametric methods are not appropriate for questionnaire data.

Keywords lighting, acceptance, questionnaires, response styles, level of measurement

ISBN (printed) 978-952-60-6065-1 **ISBN (pdf)** 978-952-60-6066-8

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2015

Pages 156 **urn** http://urn.fi/URN:ISBN:978-952-60-6066-8

A! Aalto-yliopisto

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tiivistelmä

Tekijä
Mikko Hyvärinen

Väitöskirjan nimi
Metodologisia kysymyksiä valaistuksen hyväksymistä ja siihen liittyviä mieltymyksiä tutkittaessa

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Sähkötekniikan ja automaation laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 15/2015

Tutkimusala Valaistustekniikka

Käsikirjoituksen pvm 12.09.2014 **Väitöspäivä** 13.02.2015

Julkaisuluvan myöntämispäivä 19.12.2014 **Kieli** Englanti

☒ **Monografia** □ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Valaistuksen subjektiivisten tekijöiden, kuten mieltymysten (preference) ja hyväksymisen (acceptance) tutkimiseen tarvitaan koehenkilöiden tekemiä arvioita. Tällöin tieto kerätään kyselylomakkeiden avulla.

Väitöskirjassa tehdään kirjallisuusselvitys subjektiivisen valaistustutkimuksen metodologiasta. Selvitys jatkuu kyselylomaketutkimukseen yleensä, ja työssä käydään läpi kyselylomakkeella tehdyn tutkimuksen tuloksiin ja tutkimustiedon ominaisuuksiin vaikuttavia tekijöitä. Kirjallisuusselvitys kattaa myös vastaustyylistä johtuvat virhelähteet valaistuksen hyväksymisen ja mieltymyksien tutkimisen kannalta.

Väitöskirjassa luodaan katsaus myös useisiin kyselytutkimuksen tulosten analysointiin liittyvään kiistaan: nollahypoteesin merkittävyyteen, merkittävyyskriteerin muuttaminen usean vertailun vuoksi, sekä parametristen tilastollisten menetelmien käyttäminen kyselylomaketiedon analysointiin. Kiistojen osapuolien käyttämät argumentit esitellään, sekä pohditaan sitä miksi nämä kiistat ovat kestäneet niin kauan.

Kaksikymmentä subjektiivista vuoden 2005 jälkeen julkaistua subjektiivista valaistustutkimusta analysoidaan menetelmien, sekä tulosten ja menetelmien esityksen kannalta. Tuloksena saatiin että noin puolessa analysoiduista tutkimuksista oli puutteita ennen kaikkea menetelmien, tilastollisen käsittelyn ja tulosten kuvauksessa. Tutkimusten tilastollinen voima oli myös usein varsin heikko.

Väitöstutkimuksessa tehtiin kaksi koetta, joilla selvitettiin kuinka vastausvaihtoehtojen muoto vaikuttaa tuloksiin ja niiden ominaisuuksiin, sekä verrattiin vaikuttaako valinta parameterisen ja ei-parametrisen tilastollisen menetelmän välillä kyselylomaketutkimuksen johtopäätöksiin. Tutkimuksessa havaittiin, että vastausvaihtoehtojen järjestyksen kääntäminen päinvastaiseksi ei vaikuttanut johtopäätöksiin tai tutkimustulosten tilastollisiin ominaisuuksiin, mutta kaikkien vastausvaihtoehtojen nimeäminen johti hieman suurempaan tulosten keskiarvoon verrattuna tilanteeseen jossa vain vastausalueen loppupäät oli nimetty. Sekä parametriset että ei-parametriset tilastolliset menetelmät johtivat samaan lopputulokseen, joten tehty tutkimus ei tue näkemystä jonka mukaan parametriset menetelmät eivät sovellu kyselylomaketutkimustiedon käsittelyyn.

Avainsanat valaistus, kyselylomaketutkimus, vastaustyylit

ISBN (painettu) 978-952-60-6065-1 **ISBN (pdf)** 978-952-60-6066-8

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2015

Sivumäärä 156 **urn** http://urn.fi/URN:ISBN:978-952-60-6066-8

Preface

This dissertation is a result of a long and interesting career at the Aalto University Lighting Unit (formerly the Lighting Laboratory of the Helsinki University of Technology). I am grateful for the support I have recived from my collagues there.

I am especially grateful for my dissertation supervisor Prof. Liisa Halonen, and my instructor Dr. Marjukka Puolakka for their priceless guidance and advice.

I am also very grateful for the preliminary examiners, Dr. Bert Weijiters and Dr. Teija Vainio for their very valuable comments for my dissertation, and Prof. Markku Turunen for agreeing to be my opponent.

I would like to thank also Dr. Leena Tähkämö and Dr. Eino tetri for their comments, and Dr. Pramod Bhusal, M.Sc. Muhammad Islam, M.Sc. Rupak Raj Baniya and M.Sc. Rajendra Dangol for interesting discussion during their experiments.

Lastly, I thank my family and friends, especially Anne and my late father.

SIC LUCEAT LUX

Espoo, 12th January 2015,

Mikko Hyvärinen

Contents

List of symbols and abbreviations

α	statistical significance level
β	The statistical parameter describing the propability of false negative assuming there is an effect.
δ_x	The true differene for which there is power of x
Δ_{obs}	Observed difference
$\Delta\mu$	A difference between means
χ^2	The chi-squared statistical methods
CI_x	The x percent confidence interval
D	The value of the test statistic
d	Cohen's d, a measure for the effect size
e_{acc}	the acceptable marigin of error
H_0	Null-hypothesis
H_A	The alternative hypothesis
k	Number of comparisons
N	Sample size
p	statistical p-index
$\Pr(\mathbf{X})$	Propability of X
s	sample standard deviation
ANOVA	Analysis of Variance
APA	American Psychological Association

CIE	Comission internationale de l'Eclairage, (The international Comission of Illumination)
ERS	Extreme response style
KS	Kolmogorov-Smirnov test
MRS	Middle response style
MW	Mann-Whiteney U -test
NHST	Null hypothesis significance testing
UGR	Unified Glare Rating
US	The United States (of America)
WHO	World Health Organization

1. Introduction

1.1 Background

Working knowledge of research methodology and statistical tools is necessary for anybody doing research. Lighting acceptance and preference studies focus on subjective impressions causes by lighting on human subjects. Therefore the research methods involved are similar to those on other fields studying subjective impressions and evaluations, such as psychology, social sciences or marketing research.

Many of those who work on the lighting science come from engineering background, and the formal education of engineers rarely include research methodology for subjective studies. What is emphasized in educating engineers is how to make technical measurements on some device or equipment, and how to analyse data from such measurements. A person with this background knows, for example, how an illuminance meter works and the possible causes of measurement errors associated with an illuminance measurement. A questionnaire intended to record subjective evaluations by test subject is no less a measuring instrument than the illuminance meter. Therefore, a lighting researcher needs to know at least the fundamentals on how to construct questionnaires. When this knowledge is not obtained via formal training, it is acquired through practical experience during the research work, and observing the practice of others working in the field. Sometimes, however, methodological decisions need to be evaluated more critically.

This dissertation was inspired by three observations during the more than decade the author has worked in the lighting field, and a number of questions they stimulated. The first observation was, that while the physical test set-ups used in experiments are usually reported in good detail in the lighting research literature, the reasons for using a particular approach for the used questionnaire format are usually not given. This is not intended as criticism of the choices, but leaves the question “why were the questions asked this way” open. Some use, for example, questionnaires with five response alternatives, whereas others use seven. Is there reason to suspect that one of these would be a superior choice to the other, and are the results obtained with different formats comparable? How does the way the questions are asked affect the results. There is literature on these questions in other fields, but are the results applicable for lighting research?

The second observation was, that during working with subjects, they often were concerned on how well they were doing. This observation prompted the question “how do the subjects make the decision to respond as they do, and what other factors than the stimulus and task may be involved”.

The third observation was, that instructional texts such as textbooks (e.g. [1]) make a sharp distinction between ordinal and interval data types, and the statistical methods that are appropriate with them. Strictly applied, this principle would mean that Likert-type data could not be analysed with parametric methods, and the mean is inappropriate as a measure of the central tendency. However, the practice of using parametric methods and mean with Likert-data is very widespread in the literature, and does not seem to lead to wrong conclusions.

The author was also aware of certain controversies in statistical practices, such as the use of null hypothesis significance testing, and took the opportunity to review these issues from a lighting point of view in order to understand their implications for research work. A question of interest is also “why these controversies persist?”

1.2 Scope

The scope of the present work is centred to research methodology using subjective evaluations measured by questionnaire items, primarily user acceptance studies and preference. The scope is entirely on the evaluation of subjective impressions; more technical factors, such as brightness or colour matching or task performance, are limited outside the scope of this dissertation. The focus is also mainly on the comparison of subjective ratings generated by different experimental groups, and evaluating the ratings item by item, as this is also the most common practice in the lighting field.

Although the fields of statistics, and survey research are advancing rapidly, these advances are not yet in widespread use in the practical lighting research. This has the effect of making the more advanced methods rare to encounter, and thus there are too few cases to inference patterns from. Therefore, the discussion is also limited mostly to the most commonly used and thus basic methods.

The general aim is not to criticise the existing practice but to improve awareness and hopefully stimulate discussion on the methodology.

1.3 The research questions and methods

The inspirational question for this study was: "How should subjective lighting evaluation studies be designed and carried?" This prompted three research questions:

1. *How should the questionnaire be designed?* This question was addressed by a literary review conducted on questionnaire research on other fields, and the experimental part of the present study.

2. *What potential bias sources are there when subjects respond to a questionnaire?* A literature review on questionnaires and response styles was carried out to address this question.

3. *Is it wrong to analyse questionnaire data with parametric methods?* A literature review was conducted to find answer to this question, and the experimental part addressed also this question.

The study consisted of five phases: First, two short experiments where subjects evaluated subjectively lighting settings. The subjects answered the same set of Likert-type questions but different subject groups had the response options formatted differently. The aim was two-fold; first, to find out if the way the response options are presented creates a systematic bias in the answers. The second objective was to have a "base-line" experiment at hand which can be later analysed from the basis of the literature review.

The second phase of the study was to conduct a through literature review was conducted to find out what is known on other fields about questionnaire research and the subject behaviour. The aim was to gain qualitative understanding on how the subjective lighting evaluation research should be conducted in order to maximize the data quality and avoid or reduce various biases.

Third, a literature review was conducted on several statistical issues. The crucial question was “are parametric statistical methods truly inappropriate for questionnaire data.” Other statistical controversies affecting questionnaire data were also reviewed.

The fourth phase was to analyse lighting research articles on the basis of what was learned during the literature reviews. The aim was to chart out what is considered as the methodological norm in the lighting research, how methodology was described on the research articles.

Last, the experiments conducted in the first phase were re-analysed from the basis of the literature review, to see how they could have been improved. The original aim was also to conduct another round of experiments, but it become apparent that the required sample rendered this impracticable considering the time and resources available.

1.4 Contribution

This work is entirely written by the author, who also conducted the background research and collected the material and literature used, and performed himself all the calculations and analyses presented in the work. The author also conducted the experiments described in chapter 7, and did the subsequent data analysis also himself.

The main contribution of this dissertation is that it is a pioneering work for the lighting field. While others have published reviews such as those in chapters 3, 4 and 5 in other fields, this is the first time to the author’s knowledge that these questions are reviewed especially from the point of view of lighting research.

2. State of the art

The research methodology for lighting acceptance studies has not been a subject for much research. While literature reviews on research *results* are common, broader literature reviews of research *methodology* are rare. Even when methodological reviews are presented, most commonly as an introductory part of a research article, the reviews tend to merely report what methods other authors have used rather than conduct a critical evaluation or analysis of the methods (see chapter 6).

Articles recommending a particular approach to subjective lighting research are more prevalent in the literature, as seen below. These articles are usually based on theoretical considerations and research methods from other fields, particularly psychology, or the author's own research and experience. Yet even these publications often lack a detailed analysis showing *why* the presented method or technique is better than some others.

A number of authors in the lighting field have written about scale design. Flynn et al. published in 1979 a research report titled "*A guide to methodology procedures for measuring subjective impressions in lighting [2]*" in which they addressed scaling procedures and data analysis. They used bipolar rating scale (a semantic differential scale) and multidimensional scaling as scaling procedures. In the bipolar rating scale a lighting impression was evaluated with presenting two opposite adjectives (e.g. Bright–Dim) and seven discrete unlabelled scale points between them. Flynn et al. used arithmetic means for the central tendency and analysis of variance (ANOVA) to analyse the data despite the arguably ordinal nature of the data. Although this report was published over 30 years ago, lighting acceptance studies still follow largely the same format.

Rea has also addressed scale design in lighting research. In 1982 he published an experiment on subjective scaling responses in visual performance related research [3]. Rea used a seven-point semantic differential scale influenced partly by Flynn. He found that “subjects’ task evaluation responses were based-upon the same parameters as those influencing performance but that their feeling responses were not.” Rea noted a possible lack of differentiation between feelings and task evaluations by subjects which could have serious consequences for those assuming that scaling responses are unbiased [3]. He also observed that the subjects varied in how effectively they used the scales describing some of his subjects behaving erratically and showing negative correlation in evaluating the same task. Rea argued that the subjective scales should be calibrated and suggested considering establishing standard calibration procedures for subjective scaling.

Later, in 1992, Tiller and Rea examined the patterns of semantic differential scale inter-correlations in two different data-sets [4]. They found that semantic differential scaling should not be considered as orthodox measurement, and stated that semantic differential scale experiments are meaningless themselves, but can serve as the first step in developing hypotheses. They suggested that a correlation procedure be used to make research more rigorous and using pre-experimental standards to present the subjects what is meant by e.g. “bright” or “dim” in the scale. However, as Houser and Tiller note in another paper, such a procedure could act as training the subjects in the response required, thus influencing the outcome of the experiment [5].

Fotios reviewed in 2001 the literature on apparent lamp brightness studies [6], and found that many of the studies “must be discounted due to the presence of experimental bias, experimental errors or insufficient data in the published work.” He stresses the importance of experimental design to reduce the effect of biases, by adopting a balanced design and quantified using null-condition testing [6]. Fotios discusses in a further paper the requirements for good experimental design in brightness matching tests and reviews null-condition experiment, noting a small but significant bias in null-condition brightness matching studies [7].

In 2003 Houser and Tiller compared paired comparisons and semantic differential scaling with conducting an experiment using each technique in identical lighting settings and comparing the results [5]. They note that the assumption that ratings in semantic differential scale unambiguously relate to changes in the lighting has been clearly and repeatedly shown to be false. In spite of this, they continue, many researchers use what could be argued is a discredited and flawed method. Houser and Tiller believe that more attention to experimental detail and measurement protocol are required. They suggest that it is critically important to define the stimulus and response dimensions. This can be achieved by instructing the subjects on what aspect of the luminous environment is to be rated (the stimulus), and on the precise meaning of the response formats and on how to apply them (the response). Houser and Tiller advice including an inter-correlation analysis of the semantic differential scaling responses to complement the main analyses, in order to detect possible instabilities in the scales, and complementing the data obtained by semantic differential scale with by data collected using other methods.

Fotios and Houser published in 2009 one of the very few reviews of methodology used in lighting research focusing on the perception on brightness on studies that used category rating as the principal experimental methodology [8]. They found eleven of the twenty-one studies they considered to be of dubious value, because of the lack of sufficient information to describe the methodology or the findings. A common error was missing or inadequately described statistical analysis. Fotios and Houser also found evidence of biases in categorical ratings of brightness. These biases tended to reduce the difference between ratings sufficiently to hide the differences in brightness, when the differences were small. Fotios and Houser suggest several methods to address the bias associated with the presentation order, and using an even-numbered scale (thus, one without the neutral option) to avoid the response range bias [8]. They also suggest anchoring the scale to the stimulus range by using a pre-experimental visual demonstration, whereby it is demonstrated to the subject e.g. what is meant by “very bright” or “dim” in this experiment.

Atli and Fotios studied in 2011 whether the number of scale response points affect the results in lighting context. They found that with response ranges of 5, 6, 7 or 8 points the different scale formats do not lead to significant differences in the central tendency, but will lead to different distribution profiles [9]. They also suggested that omitting the neutral option in the scale does not affect the conclusion drawn from the data .

In conclusion, reviews of subjective lighting research methodology are comparatively rare, and the number of authors writing about methodology issues on the lighting field is small. The literature is mostly about the design, use and analysis of rating formats. Virtually all the authors stress the importance of good experimental design. The literature on methodology ranges decades, but it is unclear how well it is known among the lighting researchers, and to what extent this discourse has affected the lighting research practice.

3. Studying acceptance and preference

Artificial lighting is built mostly for humans. Therefore, a great part of lighting research must be conducted with methods using humans as subjects. Some aspects of lighting, such as the perceived brightness, visibility or visual performance can be evaluated with some more or less objective measurement procedure, e.g. brightness or colour matching, visual acuity, or speed and error rate in a task. Other aspects, such as impression, appearance, preference or acceptability, are more subjective. It is said that beauty is in the eye of the beholder. However, acceptance studies need to reach such mental values and somehow translate them into numbers. This can only be accomplished by having the subjects to evaluate and report what they see, and what they feel about what they see. The questionnaire is thus a very common data collection tool in lighting research.

In 1966 Oppenheim made an observation in the preface of his book "Questionnaire Design, Interviewing and Attitude Measurement [10]" that is still very apt:

> "The world is full of well meaning people who believe that anyone who can write plain English, and has a modicum of common sense, can design a good questionnaire"

A questionnaire is often naively treated as a measurement, in which the rating measures the mental impression of the stimulus on the subject. It must not, however, be assumed that the subjects will evaluate the intended stimulus in the intended way [5]. It is, however, impossible to know the real inner meaning of every answer the subject gives, so the data must just be accepted at the face value [11, 12]. The ideal questionnaire is clear and unambiguous, and yields data that is useful, usable, valid and reliable. To maximize these qualities a fair amount of thought should be spent on constructing the questionnaire. The formation of a questionnaire requires a clear definition of the issue under consideration, and the related concepts involved [11]. An attractive option is to see from the literature what others have used. This has its advantages and disadvantages; on the one hand it makes it easier to compare the results obtained with those of others, but on the other hand it risks using and perpetuating un-optimal methods.

3.1 Acceptance and preference as attitudes

To be useful, a lighting study must have value in predicting how the same set of subjects would behave if the experiment was repeated. Therefore, the measured quantities need to be relatively unchanging over time. Psychophysics studies usually subjective judgements of stimuli that can be objectively measured on physical scales [13]. In psychophysical investigations of lighting, it has been discovered that observers with normal vision are uniform enough that a standard observer can be defined [14]. While appearance and preference are obviously not physically measurable, the fact that a standard observer can be defined for psychophysical purposes suggests that differences in preference or acceptance are not due to different psychophysical responses to the lighting stimulus. Lighting acceptance and preference can thus be considered attitudes. Attitude has been defined as "a summary evaluation of a psychological object captured in such attribute dimensions as good–bad, harmful–beneficial, pleasant–unpleasant, and likeable–dislikeable [15]." Neurological evidence suggests that evaluative judgements differ in important ways from non-evaluative judgements [15]. Evaluation is thus a critical component of attitudes, and lighting acceptance and preference contain evaluations about various aspects of lighting.

Because very little has been written on the subject from explicitly lighting point of view, most of the literature consulted below is from other fields, primarily from social sciences and psychology. This is justified, as a preference can be considered as an attitude. Nevertheless, answering a question on a lighting preference study (e.g. "In which lighting the sample objects look more natural, the left or the right") is very different from answering a question about attitudes to a complex political or social issue with ethical or moral implications. This may in fact make the responses less dependent on the desire to appear “good”.

An attitude can and does change over time, but this is assumed to happen relatively slowly. A momentary evaluation is not an attitude, but the more stable pattern emerging from these evaluations may be considered as one. It is often assumed that we have only one attitude toward any given object or issue at any given time. This may, however be too simplistic view. According to Wilson, when attitudes change, the new attitude overrides rather than replaces the old attitude [16]. Thus, people can simultaneously hold two different attitudes toward a given object in the same context, one attitude implicit or habitual, the other explicit. Over time, one attitude will probably win, usually the explicit one fading out unless it is “exercised” by use. Attitudes depend also on the context, and some apparent discrepancies between attitudes and behaviour may reflect the presence of multiple context–dependent attitudes [15], e.g. a preference reported in a questionnaire may not always reflect the purchasing choices the subject makes. Very many factors may influence the momentary preference; for example, it has been found that women’s colour preference (for cloth samples) may be influenced by such factors as the room temperature [17, 18], the intensity of previous light exposure [19], or even the phase of their menstrual cycle [17].

Acceptance and preference are thus considered as attitudes about aspects of lighting, which will remain constant long enough to be of value when predicting how people will react to the particular aspect.

3.2 The process of responding

The act of responding to an item in a questionnaire involves several cognitive steps. This requires a substantial amount of cognitive work and provides many ways in which the process can deviate from what the researcher intended. There are several different models of how people generate responses to questions, but the most commonly identified stages of the response process are [20, 21]:

1. *Comprehension* – The subjects must understand the task they are given, interpret the question and identify information sought.

2. *Retrieval* – They must search their memories for relevant information.

terpret it. For this they must understand the stimulus question, the response format, and how the two relate.

2. *Observation* – They must evaluate the visual scene from the point of the view of the question.

3. *Retrieval* – The subjects search their memories for additional information, including expectations.

4. *Judgement* – They have make a judgement integrating the observation and memories.

5. *Response selection* – The judgement is translated into the response scale.

6. *Response reporting* – The subjects will edit their response for consistency, acceptability, or other criteria and report the final response.

It should not be assumed that the response process is always conscious and deliberative. On the contrary, all the steps might happen quickly and automatically [20], and may not be complete sequential as the previous questions and experience affects already the comprehension step.

Each of the steps involves different biases and potential sources of error. The comprehension stage is vulnerable to misunderstandings and differences in interpreting the task or question. The observation stage may be biased, in addition to misunderstood task, from complex or conflicting visual setting, where the attention is attracted by something else than the object that is to be observed. The retrieval stage is biased by the expectations and what the subjects considers "usual". The judgement itself may be influenced by many things, including the setting, the motivation and the mood of the subject. The response selection stage is influenced by the structure of the response format; how easy is it to interpret and use, and how well it allows the subject to express his opinion. The response reporting stage may be affected by the need to provide a socially desirable or "safe" answer, and the subject's concern on how he is performing and is he being consistent.

3.3 Questionnaire types

Every test, questionnaire, instrument, interview protocol, scale, and so on has a stimulus component and a response component and a context component [22]. The stimulus component defines what is being measured. The response component obtains the information being sought, and encodes it into some analysable format. The choice of the response format depends on many factors, the foremost being the overall goals of the research, the experimental apparatus and setting, and the method which will be used to analyse the results. The context of the question will also influence how the question and the response options are interpreted; the same question may produce different answers in different contexts, even if the stimulus may be the same. For example, simple things like the questions preceding an item may affect the way the subjects interpret the task in hand.

3.3.1 The basic type and task

Open or close

There are two basic forms of questionnaire. The more commonly used in lighting research is the closed-ended form, which requires subjects to choose among a set of provided response alternatives. The other basic form is asking open-ended questions, in which respondents answered in their own words. These have their advantages and disadvantages.

Closed-ended questionnaires are easier to respond to, and to analyse, and they provide a ready format in which the subject can fit his opinion. This requires that the answer choices are comprehensive so that the subject can find an alternative that accurately describes his opinion, and this is difficult to assure [21]. Usually a subject needs to fit his response to the alternatives provided, which is in effect a non-linear transformation [8].

Open-formed questionnaires offer the subject a way to express his opinion in his own terms. While this is attractive from the viewpoint of avoiding distortions caused by the response format, an open-form question requires usually more effort from the subjects [23]. Forming and articulating an open answer requires significantly more cognitive effort than simply ticking a box. The subject must also be able to articulate his answer. Since each open response is unique, analysing them is more problematic and will require interpretation by the researcher. On the other hand, the information gathered this way may give more insight to the thinking process of the respondents, how they understood the question and why they responded the way they did. The lack of open questions may also signal the respondents that that individual insight is not required, and that the researcher already has all the answers [11].

Rating or ranking?

An important goal of survey research is to understand the choices people make between alternative courses of action or objects. One way to do so is to explicitly ask respondents to make choices by rank ordering a set of alternatives. Another approach is to ask people to rate each object individually, allowing the researcher to derive the rank order implied by the ratings. [21].

Researchers have typically preferred to use rating questions rather than ranking questions. However, a number of studies indicate that rankings yield higher-quality data than ratings, because respondents are more likely to make mistakes when answering rating questions, and they fail to answer an item more often than when ranking [21]. Rating is more susceptible to non-differentiation, i.e. the subject answering similarly to all the questions regardless of the content [24]. The rating task is also known to be prone to biases [8]. Ranking, however, may require greater cognitive sophistication and effort [25]. On the other hand, ranking may be more familiar because it often reflects more the actual tasks the respondents face in the real life, such as choosing what to purchase.

Ranking a large number of samples can be difficult; it could be done by having the subjects to arrange the samples in overall order, or via pairwise rankings. Both methods require the stimuli to be available to see simultaneously or closely sequentially, which more or less rules out experiments done in full scale mock-up, or real office rooms.

If a large number of samples are ranked with pairwise comparisons, a large number of comparisons are required, which makes the process cumbersome. Pairwise rankings do not always constitute a range, and transitivity may not always apply in rankings. For example, let us consider a pairwise comparison experiment where three light sources (A, B and C) are compared pairwise so that all three possible comparisons (A–B, A–C and B–C) are made. If we assume that transitivity holds, then, if lighting scene A was preferred to lighting scene B, and B to scene C, then we would expect that A was also preferred to C (see figure 3.2.) . However there is

Two alternatives (binary)

The simplest (in terms of the amount of information gathered – one bit) response format is a choice of two alternatives (Figure 3.3. a). Such choice can be, for example, a yes–no -choice ("is there glare"), a comparison between two scenes ("which scene is more natural, left or right") making the task essentially a ranking, or semantic opposites ("lighting is good–bad or adequate– inadequate"), or an agree–disagree choice ("Energy efficiency of lighting is an important factor to me when choosing lighting") .

Two alternatives with side by side estimation between two stimuli is a method much used in the lighting field. It is easy to use, and clear and much simpler for the subjects than estimating a rating [8]. This response format divides responses in clear categories, from which proportions are easy to calculate and analyse. On the other hand, such scale may be too crude for subtle differences, and the answers are qualitative only.

Paired comparison with expanded scale

Paired comparison can also be rated with a range with more than two points (Figure 3.3. b); one could, for example, ask if the subjects preferred the scene on the left-hand or the right-hand compartment with a response format, e.g. with a number of options, or a visual analogue format (see below), which the respondents can use to express how much they prefer one over the other.

Estimation

A simple way to quantify estimates is to ask directly for a numerical estimation. For example, "On a scale of 0 to 10, how bright the scene looks?" as in the Figure 3.3. c). The responses usually generate quantitative data. However, estimating may be a difficult task; many people find adjectives a more natural way to express their estimations [13].

Such question is sometimes also hard to anchor, but can be made clearer with labelling the alternatives. Judging magnitudes that do not have familiar physical units can also easily create biases [8] and requires the subject to place an observation on a range without reference to the range being used [6].

Likert-type

A Likert-type response format has a statement or a question, and several response options, typically five or seven, which the respondents use to express the degree to which they agree with the statement (e.g. "fully

a) In which scene the lighting looks more *natural?*

Left □ Right □

b) In which scene the lighting looks more *natural?*

Left □□□□□□□ Right

c) In a scale from 0 (not natural at all) to 10 (very natural), how *natural* does this scene look?

0 1 2 3 4 5 6 7 8 9 10

d) This scene looks *natural* to me

Fully disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	agree	Fully agree
□	□	□	□	□	□	□

e) This scene looks to me

Completely unnatural □□□□□□□ Completely natural

f) This scene looks to me

Completely unnatural ―――――――――― Completely natural

Figure 3.3. Examples of response formats. a) two choices, b) a paired comparison with an extended scale, c) rating with a scale, d) a Likert-type scale item, e) semantic differential and f) a visual analogue scale (measuring semantic differential here).

agree, somewhat agree, neither agree nor disagree, somewhat disagree, fully disagree") (Figure 3.3. d) [22, 26, 27, 28, 29]. The alternatives may be either verbal as in the example, or a numerical ("indicate your agreement from 1 to 7, where 1 is fully disagree and 7 is fully agree) [26]. Likert transformed these categorical answers into numbers and used the sum of these numbers as the final score, but the items are often considered individually. However, some argue that the items should not be analysed individually at all [22], while others deem it an error if they are analysed as a compound instead individually [8]. This may depend on the context and the exact way the questionnaire is used.

Likert-type response format is usually easy to understand, and asking the personal opinion serves to motivate the respondents. On the other hand, there often is only one way to agree, but several to disagree. Only the view in the statement of the item is explicitly presented to respondents, and all the alternative views are supposedly included under the "disagree" options [21]. While the item statement formulates the "agree" positions clearly, the "disagree" position is left unarticulated, which forces the disagreeing respondents to formulate it themselves. Thus, agreeing requires less cognitive effort than disagreeing. Also, the respondents disagreeing with the statement are forced to be contrary in order to represent their views [12]. These factors may make Likert-type items more vulnerable to acquiescence effects (see section 4.1), i.e. to agree to some degree regardless of the content of the questions. This is supported empirically; responses to Likert-type questions are known to tend to lean more to the positive side of the option range [26].

Semantic differential

A semantic differential format presents the respondents a pair of opposing adjectives, and the respondents are provided a way, typically five to seven discrete options, to rate their view along an axis between the positions described by the adjectives (Figure 3.3. e). For example, “Lighting here is” could be presented with a seven-point range from dim to bright.

Semantic differential formats are very popular because of their ease to use and versatility. They provide a way to ask a large number of questions in a short period of time and a limited space. The shortcomings are related to the difficulty of choosing the adjectives so that they are equally strong.

Visual analogue format

A visual analogue response format is a way to present a situation e.g. similar to estimation or semantic differential scales, but without limiting the responses to a set of discrete options. In visual analogue format the subject is presented with a line, where the subject marks his estimation (Figure 3.3. f). The results are then measured with a ruler from the line.

Visual analogue formats have been widely used in medicine in estimating pain, and they have seen some use in lighting studies, too (e.g. in [30, 31, 32]).

3.4 Factors affecting the quality of data

Alwin and Krosnick argued that the reliability of an attitude survey depends on a number of factors, including the characteristics of the population of interest, the topics assessed by the question, the design of the questions (e.g. wording, context, response format), the observational design, the mode of administering the questionnaire, and the social situation during the experiment in general [33]. The reliability of the research is thus not just a function of the measurement instrument, but is born from an interaction between the questionnaire, the subject, the designer of the questionnaire, and the person administering the questionnaire. The literature for survey design is plentiful; a researcher should have at least a passing acquittance with a good review piece on the subject.

Mode of administering the test and experimental conditions

When we make observations, the observation conditions can affect the results significantly, and care must be taken to vary only the factors that are being studied. Standardized viewing conditions are used in many industrial applications, but in research they can often vary, which can make it difficult to compare two studies with somewhat different viewing conditions.

At least six factors need to be controlled, in addition to the objects being observed [14]:

1. The spectral properties of the light source

2. The intensity of the light source

3. The angular size of the light source

4. The angle of incidence (the direction from which the light arrives at the objects)

5. The angle of viewing (the direction from which the object is viewed (angle of viewing)

6. The background

Additionally, care must be taken to maintain other, non-lighting conditions, e.g. temperature, air quality and noise level, reasonably constant. Sufficient time for brightness and chromatic adaptation must also be given before conducting the experiment.

The form of administering the questionnaire may influence the results, i.e. do the subjects fill in the form themselves or do they answer the questions from the researcher who writes them down.

A null-condition test should be used to asses the effect of any unintentional differences in the settings. The null-condition presents identical stimuli to the observer, thereby quantifying the effect of any situational biases [7]. This should be done both as a part of the preliminary pilot experiment, and as a quality control measure during the actual experiment.

Subjects' motivation

To obtain high-quality data, it is necessary that the subjects evaluate carefully enough what is asked and respond in an honest, consistent and unbiased manner without trying to second-quess what he is "supposed" to answer. This ideal is never completely realised in practice. The motivation of the subject influences greatly whether he will perform the necessary cognitive work instead just providing some responses. The behaviour of the subjects forms a continuum from thoroughly and unbiasedly performing his task (Krosnick calls this "optimizing") to responding to first reasonable option offered looking for cues for a "safe" answer, on to responding randomly or uniformly regardless of the question (Krosnick calls the latter behaviours weak and strong forms of "satisficing") [34]. The subjects may be initially motivated to perform their task diligently, but then get fatigued or bored and start satisficing.

Respondents who are uncertain about the intent of a question may ask the interviewer for clarification. The researchers are usually concerned that all respondents have the same information, which means that the experimenters are usually not allowed to provide substantive help [23]. This leaves the respondent alone to formulate a defensible response strategy, which can lead to satisficing behaviour.

The creation and maintaining the motivation for the subjects is something that must be considered when designing the experiments – an automatic "somewhat agree" to every question is hardly the kind of data the researcher was after. The motives for optimizing behaviour include desires for self-expression or self-understanding, for interpersonal response, for intellectual challenge and the gratification for successful performance, or altruism or to help getting products that suit the consumer better. The undesirable satisficing behaviour is more likely when the subjects face a more difficult cognitive task, their ability to do the task is lowered, or their motivation is lowered. [34]

Researchers often think that it is best to tell the subjects as little as possible about the experiment in order to avoid influencing them. It may improve the quality of the data obtained if attention is paid on the motivation of the subjects. The subjects can be motivated by stressing the importance of giving the task their full attention. For example, the subjects might be told that "the results will be used in developing lighting recommendations, so it is important that you concentrate on the task." Or, "The differences between the scenes are small, so it is important that you look carefully." Even when the objective of the research is to study the immediate impressions, the subjects should not be instructed in a way that seems to downplay the importance of their answers. Thus, it is better to say: "We are interested your first impressions of the lighting" than "just your first impressions." The persons administering the questionnaire should be trained so that they understand the significance of the motivation.

Respondents' preference for scale formats need to be considered. For example, if a scale is too difficult to use, or too crude to allow respondents to express themselves, the respondents may become frustrated and demotivated, and the quality of their responses may decrease [35]. On the other hand, too simple and repetitive scale may also prompt mechanical responding behaviour where the subject just ticks the responses without really thinking the question, so it may be good that the task is complex enough –but not too complex– to keep the subject focused.

Blinded experiments

In clinical medical research, it is an important factor to make the trial double-blind, i.e. neither the subject nor the person who conducts the experiment in practice, knows at the time if the subject is in the control group or in the test group. In the field of medicine, it has been shown that studies with lacking or inadequate blinding show more positive treatment effect than those where blinding has been done appropriately [36, 37, 38]. The effect has also been demonstrated in medical research conducted on animals [38]. Thus there are empirically proven risks of bias toward inflating the results if adequate blinding is not used. This may seem to be challenging the integrity of researchers, but it is usually not a case of fraud; these biases happen unconsciously [39]. The bias can arise from unconscious bias during assessing the results, or the researcher unintentionally signalling his expectations during the experiment. The subjects can and do look for unconscious cues for the "desired" answer from the person administering the experiment.

Lighting experiments are usually designed so that there is no special control group. It is a common practise to keep the subjects unaware what lighting they are seeing, but the literature mentions practically never a test procedure where the person conducting the experiment was also unaware of the used lighting. Often the researcher-side blind experiment is impossible to realise, as the lighting condition being used may be visually obvious to someone who knows what the alternatives are. When the setting is built so that this is not an issue, the experimenter side of the blinded procedure is easy enough to realise; all that is needed is that one person sets the lighting condition, and another administers the experiment to the subject.

Wording and translations

The exact wording of the questions have a great influence. The questions should be inherently obvious to their meaning. It should be remembered that a layman will not understand a technical term (e.g. glare) in the same way than an expert. Each statement should be as short and simple as possible [40]. The language should be familiar to the subjects; the use of jargon, unusual words, acronyms and abbreviations should be avoided where possible [11, 40]. Consideration should be given to whether the words chosen have an alternative meaning, and the researcher should beware of implicit assumptions in question wording [11].

Questions should generally be worded so that they are affirmative, i.e. do not include negations [11, 20]. If reverses are needed, it is better to use antonyms as opposites than negations (i.e. if a reverse is needed for “Bright” it is better to use “Dim” than “Not Bright.” However, care must be taken to avoid leading questions, where the wording suggests the appropriate or socially desirable answer [11]. See also the discussion on acquiescence in section 4.1.

Double-barrelled questions, i.e. ones where a single question or statement includes two (or more) assertions must be avoided [40]. For example, a Likert-type evaluation might include the stimulus statement “More innovative street lighting solutions should be used for improved safety, energy efficiency and appearance of the outdoor environment”, which actually contains three or four claims. A respondent might agree that safety must be improved, but be indifferent to the other two reasons.

Respondents presume that researchers provide less important background information first and then present more significant foreground information later [21]. The reference period may also influence how a question is interpreted. For example, a question about how often the respondent has been angry in the past year is typically interpreted as referring to more intense emotions than a question that refers to the past week [23].

Translations cause special problems. Not only is it often difficult to come with translations with exactly similar content, but the roots of the words can make the "dictionary-equivalent" words to have different connotations. The strength of qualifying words and phrases can vary greatly; In WHO's work on developing the WHO Quality of Life measure, Szabo described a study to establish the equivalence between response categories in nine regions of the world. He found, for example, that "quite often" in England appeared equivalent to "often" in India, "from time to time" in Zambia, "sometimes" in Melbourne and Seattle, "now and then" in the Netherlands, and "usually" in Zagreb [13].

The range of choices can influence how an individual item, even if described with word, is interpreted. It is, for example, possible that the option "agree" is understood differently when the "agree" represents the end anchor of the rating scale, compared with when the range of possible responses also includes the option "fully agree" [41].

Varying response formats

Varying response formats will reduce monotony and force the subject concentrate more on the responses he gives. On the other hand, variation will also make the questionnaire more confusing.

Using similar response formats and anchors makes it easier for the respondents to complete the questionnaire, because the standardized format requires less cognitive processing to answer. However, this may cause some of the co-variation observed to be due to the similar response formats, rather than the content of the items [20].

Labelling

How the response options are labelled can effect how the respondents interpret them [21, 42, 43]. The labels, e.g. numbers are often selected arbitrarily (e.g. a 7-point range can be labelled from 0 to 6, 1 to 7, or -3 to +3), but the respondents may presume that how the options are presented contains information on the intended meaning or "the correct" responses.

Some ways to provide the labelling is shown in the Figure 3.4. The cases are otherwise all similar, i.e. a semantic differential format with seven discrete response options, including a neutral one.

The format in the Fig 3.4. a) shows an unlabelled range from 1 to 7. Such range may confusingly appear as unipolar, i.e. from "little to much", whereas a range from -3 to +3 as in the 3.4. b) is more clearly a bipolar range, from “very negative” to “neutral” to “very positive”. The version in the Figure 3.4. b) is also more clear as to where the neutral option is located; the 1–7 range requires some counting to know that the “4” signifies the neutral. The format in the Figure 3.4. c) is very close to that in the 3.4. a), except the options are labelled with letters instead of numbers. The difference, if any, is that numbers appear more naturally as points on a continuous range, whereas letters could be seen more clearly as categories.

The response option format in the Figure 3.4. d) omits the symbols that differentiate the options, relying on the visual position of each option to signal their value on the range. This avoids any signals associated with numerical labelling. While such scale reduces the risk of influencing the respondents inadvertently, it also misses the opportunity to clarify or reinforce the meaning of the response options with suitable labels. The end-point labels have also been placed to the actual end-points rather than the text. The format in the Figure 3.4. e) is otherwise similar, but the neutral option is marked with printing the mark in bold. This may make the response options quicker and easier to understand, and provide a quick way to make a good–bad -judgement, but formatting one option differently from the others may be seen by some as a cue for the “correct” answer.

The response formats in the Figures 3.4. f) and g) provide verbal labels for every option. Both need more reading than the formats discussed previously, which may prompt some subjects to stop at the first defensible option. On the other hand, the response alternatives are clearly marked. Such formats take more space than those above, which may make a questionnaire of similar number of questions appear longer than if done with a more compact response formats. The horizontal tables in the Figure 3.4. f) also have a limit on how wide the alternative box is, which leaves description texts hyphenated and arraigned in narrow columns.

One alternative is to label the response options with graphics as in the Figure 3.4. h). While this is an intuitive way to label the alternatives, it often limits severely the wording questions that can be asked with such formats; The one presented suits at most to a survey where every question asks “how satisfied/happy” the respondent is with a particular aspects. Even then, the faces as such describe a degree of joy or sadness, not strictly speaking content/discontent or like/dislike.

Garland studied the effect of labelling in a semantic differential survey [44]. He found that labelling did not create significant differences in

a) How *satisfied* are you with the current lighting? (1=very unsatisfied, 7=very satisfied)

1 2 3 4 5 6 7

b) How *satisfied* are you with the current lighting? (-3=very unsatisfied, 3=very satisfied)

-3 -2 -1 0 +1 +2 +3

c) How *satisfied* are you with the current lighting? (a=very unsatisfied, g=very satisfied)

a b c d e f g

d) How *satisfied* are you with the current lighting?

very unsatisfied O O O O O O O very satisfied

e) How *satisfied* are you with the current lighting?

very unsatisfied O O O **O** O O O very satisfied

f) How *satisfied* are you with the current lighting?

very unsatisfied	unsatisfied	somewhat unsatisfied	neither satisfied nor unsatisfied	somewhat satisfied	satisfied	very satisfied
□	□	□	□	□	□	□

g) How *satisfied* are you with the current lighting?

a) very unsatisfied
b) unsatisfied
c) somewhat unsatisfied
d) neither satisfied nor unsatisfied
e) somewhat satisfied
f) satisfied
g) very satisfied

h) How *satisfied* are you with the current lighting?

😫 ☹ 🙁 😐 🙂 ☺ 😃

Figure 3.4. Examples of labelling.

the ratings, but participants clearly preferred to use the labelled form. It should be noted, that in this study the format where all the points were verbally labelled was judged not only the easiest to understand and most preferred, but easiest to complete as well, even if there are more information to read! When he asked why they preferred the labelled variant, he found that the majority of participants wanted definite options to aid them in making a decision. The presence of verbal cues on the labelled semantic differential was seen as offering reassurance and making the task more or less self-explanatory. Precise answers seemed to be important to respondents, and the verbal tags were seen as aiding precision. However, those of the respondents who were used to working with numbers often preferred the form with numerical labels. [44]

This suggests either that rating-scale points should be labelled with words. Krosnick suggests that the numbers should be carefully chosen to reinforce the meanings of the words, and the items checked to avoid conflicting meanings between the wording and the response options [21]. Schaeffer and Presser claimed, however, that verbal and numeric labels appear to have separate effects that do not interact [23]. In any case, the researcher should ask himself whether the numbers on the response options are really something the subjects need to see, or are they just used by himself to analyse the data.

Some suggest providing verbal label to the extreme points and to the mid-point (if the response options include one) [20, 45]. However, if the response format seem to emphasize some, but not all, response options by labelling them, the subjects may take the labels as cues for the "desirable" answer. The labelled options attract more responses also because they need less cognitive effort to process [43]. In this view it may be better to label either all the options, or just the end-points, rather than the middle point in addition to the end-point.

Labelling may also affect the way the responses are distributed. When an option is labelled, it is more clearly understandable and therefore accessible to the subjects. Therefore, when only the end-points are labelled, the intermediate options may be seen as less accessible to the respondents, and thus access less responses than if the intermediate options were also labelled. On the other hand, if the stimulus question prompts agreement or disagreement, labelled options reinforce the sense of disagreeing, which may result to results leaning to more positive side due to acquiescence effects (see Section 4.1). [43]

Anchoring

Rea argued that subjective scales should be calibrated and suggested considering establishing standard calibration procedures for subjective scaling [3]. Anchoring has been suggested as a way to demonstrate how the response range applies to the stimulus range by using a pre-experimental visual demonstration [5, 8, 46], i.e. what is meant by bright or dim in the context of the experiment.

However, this may risk influencing the subjects in other ways. It may also be impossible to do – illustrating brightness may be easy, but it is much more difficult to anchor such concepts as e.g. naturalness or pleasantness, let alone the like – dislike axis. Anchoring is a viable approach for studies in which all the interesting questions can be anchored, but it may be confusing for the subjects when only some of the questions are.

The number and order of the questions

The optimal number of question is as few as possible, but sufficiently many; every question must be absolutely necessary and serve a crucial function in eliciting the data required. The researcher should seek the minimum of information necessary, as the respondent's time is precious, and asking unnecessary and irrelevant questions will increase the questionnaire completion time. Too many questions may have an adverse effect on the respondent's motivation. On the other hand, repeat questions, i.e. a question with the same objective, but worded differently, are known to increase the reliability of the data. [11]

Background questions (e.g. age, sex, whether the subjects wear glasses) are often asked first, either just because it is customary to do so, or to ease the respondent into the questionnaire and increase their confidence. However, seeking demographic details can be perceived as threatening in some circumstances, and starting with a tranche of personal questions may well dissipate the enthusiasm of the subjects. In this view it could be best to ask these questions at the end of the session. [11]

Questions on a similar theme should be grouped together for cohesion, and questions should flow smoothly from one topic to another. Questions on each topic should be completed before compiling questions for the next section, and at the questions should develop in a manner that approximates to the respondent's view of reasonableness and logic. [11]

Krosnick showed that the order in which information is provided in the stem of a question is sometimes viewed as providing information about the importance or value the researcher attaches to each piece of information. Specifically, respondents presume that researchers provide less important background information first and then present more significant foreground information later. [21]

Questions tend to activate beliefs that are consistent with the way in which the item is stated [47]. For example, if asked about how natural a lighting scene looks, the respondents will look for the ways in which the lighting is natural, but if the question was about how unnatural the lighting looks, the respondents would look for factors making the scene unnatural. These beliefs may linger on when the subject evaluates the subsequent items; if the subjects are asked first about a certain aspect of lighting, this primes them to think the lighting from that aspect which may affect the results of the latter items [20]. For example, a negative bias may result if the subjects are first asked several items about the negative aspects of lighting or items making negative assertions or worded negatively (e.g. "How bad is the glare?", "How would you improve this lighting?" or "The lighting here is too dim"), and then asked a question about how satisfied they are with the lighting in general.

Subjects often assume that questions that are located closely together are related [20], so if the purpose of the questionnaire is to find correlations it might be good to have some neutral questions between the items which are especially interesting.

Subjects are constantly making small adjustments to their internal reference during an experiment [8]. The respondent's evaluations are affected by both the present and previous stimuli. Thus, the response to a particular stimulus tends to be biased toward the response on the previous one [8]. When multiple judgements are made by a respondent using the same scale, respondents use their initial ratings to anchor the response range. This is yet another way for the earlier questions to influence the responses to those coming later [20].

The simplest method to avoid the order bias is to mix the order of the stimuli or judgements so that, while the effect of the previous stimulus is not eliminated in each individual estimation, the overall bias across all the subjects is cancelled [8].

Number of response option alternatives

The choice of the number of response categories is a compromise between the increasing discrimination potentially available with more categories, and the limited capacity of respondents to make finer distinctions reliably [23]. The question of the number of response option alternatives can be approached from several perspectives. From the point of view of information theory, the more there are options, the more bits of information is gathered [48]. A statistician will probably consider what difference the number of categories makes to the statistical characteristics of the data. A third approach is to consider the question from the cognitive point of view; how the subjects will understand the response range.

The number of response option alternatives can affect the characteristics of the gathered data [26]. The most commonly seen practise is to use a five- or seven-point response range [4, 5, 26, 35], which are the ones usually portrayed in textbooks as well [26]. A range of seven response options is recommended for semantic differential response formats in the lighting field, too [4, 5]. However, 10- or 11-point ranges are used often as well. It is also possible to use a continuous range of responses, such as in the visual analogue scale. Technically, especially in an questionnaire conducted with a slider in a computer screen, it may be in fact a range with a large number of discrete options (e.g. 256), which in practise behaves like if it was continuous.

Too many categories may go beyond a respondent's ability to distinguish among categories [48]. Even a pen-and paper questionnaire asking the ratings as, e.g., a percentage (0 % – 100 %) is providing a far finer scale than the human estimation requires – Miller suggested in 1956 that on average the human mind has a span of apprehension capable of distinguishing about seven different items [49]. If a large number of categories are presented, the respondents tend to use multiples of 5, 10 or 25, so the actual number of response categories used may be smaller than what would appear [23]. Wyuts et al. compared the visual analogue format with a four-point discrete format, and found that the inter-rater agreement decreased considerably with increased freedom of judgement, recommending that the four-point format be preferred to the visual analogue scale [50].

Theoretically, there are several mechanisms via which the number of response options can affect the data characteristics. The arithmetic properties of response range may result in a slightly different mean scores and skewness when the majority of the true opinions were on one side of the midpoint [26]. A broader range of response options provide more options for the respondents to choose, which in turn may result in a greater spread of the data and thus a larger variance and a negative kurtosis (i.e. a less clearly peaked distribution) [26].

The optimal number of response alternatives depends on the mode of administering the questionnaire. For example, if the alternatives are read aloud by the experimenter, who also marks the responses down, reading the full list of ten or more alternatives for each question becomes quickly cumbersome. If each response option is verbally labelled and a wide range of response options are provided, it will become difficult to come up with satisfactory qualifiers which clearly separates the options in a steady progression. For this reason, questionnaires using a larger number of response options typically lean more on expressing the options as numbers for which the precise meaning has not been defined [26].

There are many studies on how the response format affects the reliability and validity, with the general result that reliability is largely independent of the number of response categories, at least if there are at least five to seven options provided [40, 35]. The reliability is somewhat lower when there are very few (less than five), but only trivial gain in reliability with more than seven categories [35, 51]. Validity also has been shown to increase with increasing number of response options, but with only small or trivial gains after about seven response options [35].

Preston and Colman studied the effect of response format with data from student restaurant users survey. They used response ranges with 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11 options, and a 0 to 100 rating task where no options were displayed [35]. They also asked the students to rate the rating formats in terms of "ease of use", "quick to use", and "allowed you to express your feelings adequately" to find out the respondents preference for the response formats. Their results for reliability and validity were similar to those described above. The three rated the easiest to use were the ones with 5, 10 or 7 options. The "quick" to use followed linearly the reverse number of options provided, and the higher the number of options the better were the format rated in terms of "allowed to express your feelings adequately." Taking all three respondent preference ratings into account, scales with two, three, or four response categories were least preferred, and scales with 10, 9, and 7 were most preferred [35].

Less attention has been paid to how the response format affects data characteristics such as mean, variance skewness or kurtosis [26]. Dawes studied the effect of number of response options in an marketing research experiment using a telephone survey, and found that the relative (i.e. re-scaled) mean scores were slightly lower with a 10-point response range than with a 5- or 7-point one [26]. He found no effect for variance, skewness or kurtosis.

Atli and Fotios have studied this with asking students evaluate the indoor environment of their lecture room [9]. They found that with response ranges of 5 ,6, 7 or 8 points the different response formats do not lead to significant differences in the central tendency, but will lead to different distribution profiles, which is a reverse result from the Dawes study mentioned above.

In general, the effects produced by differing number of response option is small, so the common practice of using a seven-point scale seems justified [9, 26, 35].

Response option order

The order on which the response alternatives are presented to the subjects have been shown to affect the results, but it is not clear when these effects occur and what their direction is likely to be [21, 25]. There are two types of response order effects: recency effects and primacy effects. A recency effect occurs when response options near the end of a list of alternative responses are more likely to be selected, and the with primacy effects this probability increases when the response option is near the beginning of the list of alternative responses [25]. Many experiments designed to examine response-order effects found none, some studies identified primacy effects and other studies found recency effects [21, 25, 41].

Primacy effect may be explained by respondents who are weak satisficers and simply choose the first reasonable response alternative [21]. Exactly which alternative is most likely to be chosen depends on how the response choices are presented.

If the options on a rating scale are presented visually, primacy effects are more likely [21, 25]. According to Krosnick, this occurs for two main reasons: First, the items presented early may establish a standard of comparison guiding the interpretation of later items, and thus, early items may have special significance in subsequent judgements. Second, the items presented early are likely to be subjected to deeper cognitive processing; by the time a respondent considers the later alternatives, his or her mind is likely to be cluttered with thoughts about previous alternatives that inhibit extensive consideration of later ones [25]. Primacy effects can also be caused by satisficing; the respondents proceed along the list and choose the first acceptable alternative instead going through the entire list and choosing the optimal one [25].

However, when the response options are read aloud to the subjects, recency effects are more likely [21, 25]. In such setting, the subjects are not given the opportunity for extensive processing of the first alternatives offered but have to concentrate on listening the following ones. Thus, the presentation of the second alternative forces the subject to stop processing the first one. Under these circumstances, the subjects have the opportunity to carefully consider only the items at the end of the list, since interviewers usually pause after reading them [25]. It is also easier to remember the first and last-mentioned items than the middle ones when the list of response alternatives is not visually seen by the subjects [25].

Response format effects may also be due to different interpretations where the perceived neutral point is [41]. For example, if the subjects were asked to evaluate the question “The lighting here is too bright” with a Likert-type response format there are two strategies to form the neutral; some might consider the “fully disagree” position to require too dim lighting, regarding the range as bi-polar one, whereas to others it is sufficient that the lighting is “just right” to “fully disagree” with the statement that it is too bright, in effect considering the range as unipolar. Hofmans found that for those that consider the neutral to be in the middle of the range the presentation order had no effect on the perception of the verbal qualifiers, but those for whom the neutral was the extreme option (fully disagree) showed more agreement when the degree of agreement was increasing compared to when they were decreasing as one proceeded along the response options [41].

An obvious strategy to counter biases arising from response option order is to use two versions of the instrument with reversed response option orders. Approximately half of the subjects would be given each version, and the presence or size of the response option order bias can be estimated by comparing the results from each version. It is rare, however, to see mentions of a procedure such as this in lighting literature.

Should there be a neutral point?

Both Likert scales and semantic differentials usually use odd number of steps, thus having a neutral mid-point. The neutral option can mark true neutral, indifference, or ambivalence, depending on the context. Therefore, the definition of the mid-point may affect the meaning of all the other points as well [23].

There is some debate whether or not the respondents should be allowed a neutral opinion, though it must be accepted that in some cases the respondents may be genuinely neutral about the questions [9, 11, 48]. A mid-point provides an defensible alternative for people with no defined attitude towards the issue under discussion [48]. However, at least with the semantic differential formats, the neutral may be the only option where the situation is good or suitable. For example, if the task is to evaluate the hue of the lamp colour with a "too cold – too warm" -range, the neutral point is the only Goldilocks "just right, this I like" -option; all the others would indicate dissatisfaction of varying degree.

Omitting the neutral point may make the mental "distance" between the response options uneven; consider the situation in the figure 3.5. What does the removal of the neutral option do to how the response option range is perceived? It could be argued that, at least in the cases where the endpoints are labelled, the rest of the scale would merely "stretch" between the fixed endpoints staying more or less uniform, and this is what the statistical analysis assumes if the options are coded into numbers and the result analysed as data was of metric interval nature. In the case of labelled responses, the distance between "somewhat disagree" and "somewhat agree" is two steps in the version with the neutral option included and just one in the one without the neutral option, while the labels for the other values remains fixed. It is less clear if the mental values of the options as understood by the subjects remains as fixed.

The traditional view suggests that the qualitative results are similar for a response range with and one without the neutral point, because if the respondents are truly neutral, then they will randomly choose one or the other side of the issue thus not creating a bias in the overall results [52]. It is unclear, however, whether people actually answer questions randomly [23].

Respondents who experience more ambivalence will more strongly prefer the extreme alternative when using a response format with even number of categories [52]. Including a neutral position will systematically affect attitude response distributions when attitude toward an object is ambivalent [52]. Less intense respondents are more affected by the presence or absence of a middle response category than respondents that feel strongly about the attitude [48].

Atli and Fotios found that omitting the neutral does not affect the con-

This scene looks to me

Completely unnatural □□□□□□□ Completely natural

This scene looks to me

Completely unnatural □□□□□□ Completely natural

This scene looks *natural* to me

Fully disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Fully agree
□	□	□	□	□	□	□

This scene looks *natural* to me

Fully disagree	Disagree	Somewhat disagree	Somewhat agree	Agree	Fully agree
□	□	□	□	□	□

Figure 3.5. Omitting the neutral response option; does the mental "distance" between the options change if the neutral option is omitted?

clusions drawn from the data; explicitly offering the middle position significantly increases the size of that category, but does not otherwise affect the middle position [9]. They suggest that it is better to omit the neutral option if the interesting question is in which direction the subjects are leaning, but include it when the objective is to sort out those with more definitive options. In another study, Fotios suggests that rating formats should avoid an obvious centre to avoid the contraction bias [8]. The results of Preston and Colman on their respondent preference survey shows no systematic effect for odd/even number of response options [35]. Moore et al. compared five- and six-point ranges, and found that the six-point range (i.e. without the neutral option) produced more consistent behaviour [48].

In conclusion, omitting the neutral option may improve the quality of the data, but the decision needs to be made only after carefully considering the experiment, task, response format and question.

Should a "no opinion" option be provided?

Some respondents may feel a pressure to always offer an opinion when they in fact have none [33, 53, 54, 55]. The subject may think that responding with "no opinion" may make them seem ignorant [53], or that make them feel that they not performing their task properly [55]. The subjects with no clear opinion or who have not thought about the question may cope by responding randomly [33, 55]. Worse, a satisficing subject may not respond completely randomly, but with any option which seem defensible [34], or as a response set where every question gets the same answer regardless of the context of the question [11]. The ultimate distribution of the data generated by these response strategies is strongly influenced by the structure of the questionnaire, but in any case the data does not answer the question that was asked. Krosnick et al. found that in a political attitude survey the frequency of responding "no opinion" increased when respondents voted secretly, when questions were asked late in a survey, and when respondents devoted little effort to answering question [55]. This result would imply that the "no opinion" is often not an accurate statement, but satisficing or acquiescence.

On the other hand, some subjects may have an opinion but be uncertain about expressing it. All questionnaires represent a potential intrusion into the respondent's private mental life, therefore care should be exercised to avoid embarrassment and promote the confidence to give an honest response [11]. If the subject is apprehensive about expressing his true opinion, the neutral or "no opinion" options may appear as the "safe" alternative [12, 55]. Thus, the "no opinion" may attract also respondents who actually have an opinion, and who would have reported their opinion had the "no opinion" option not been offered [55]. The "no opinion" may be a too easy way to avoid the cognitive work needed to answer the questions [55]. Many subjects will report their opinion the best they can if not offered an explicit possibility to not to [24, 55]. Thus, a researcher offering the "no opinion" option may rob himself of meaningful data by effectively reducing his sample size.

The evidence of the effects of providing the "no opinion" on the quality of the survey data is mixed and even conflicting [55]. According to many studies, offering a no-opinion response option increased the proportion of respondents who declined to report each item [55]. However, there is no clear evidence that offering the "no answer" option increases the quality of the data [55].

The ease to respond with "no opinion" depends on how the questionnaire is administered. If the subjects fill in the form themselves, they can always respond with "no opinion" by simply not answering a question. The researcher cannot know whether the subject did not have an opinion, did not want to divulge it or simply skipped a line by mistake. But when the research set-up was such that the researcher reads the questions and marks the answers after verbal responses, not answering becomes harder as it will require challenging the experimenter.

A high proportion of “no opinion” responses may be more a signal of poor question design than widespread lack of opinion [33, 55, 56]. For example, some "don’t know" answers may occur because respondents are not sure how to interpret the question stem or the response choices and might be reluctant to make a decision in this regard [55]. Other obvious case is a question which has no relevance for some respondents.

It should be stressed that the true neutral opinion is different from the no opinion and non-disclosed opinion. In an ordinal or interval scale the neutral mid-point represents a clear location along the response range, but “no answer” is not a point on the range, just missing data.

Consider a hypothetical question (using a semantic difference scale) on how bright the presented lighting is. The following are some cases of what the subject might be thinking while checking the neutral or no opinion box:

- *“I think the lighting here is neither too bright nor too dim*” is a case where the neutral option on the scale is a true and accurate assessment of the opinion.

- *“It’s a bit dim to read but that light there is too bright and glaring so which should I check?*” is a situation where “no opinion” is caused by complex or ambiguous setting.

- *“I don’t understand the question, does it mean the desk or the entire room?*” is a “no opinion” situation caused by unclear phrasing of the question or unclear setting where the subject has difficulties in understanding how the question applies to the situation.

- “*Whatever*” is a “no opinion” case caused by the lack of motivation by the subject.

- *“I don’t want to say because I may make fool of myself if my responses are not right*” is not a case of no opinion, but of the respondent being unwilling to express his opinion, even if he has one (and he probably does).

Therefore, if the “no opinion” option is used, it might be necessary to include *both* the neutral option and the “no answer” option. If there is no explicit “no opinion” option, there is a risk that the neutral, if present, will be used as one.

4. Response styles

A response bias is a systematic tendency to answer questionnaire items on some other basis than an accurate answer to the question [48, 57]. A consistent bias by a single responder across items and method is called a response style [58, 57]. Three response styles are prominent in the literature: acquiescence (tendency to agree), extreme response bias (tendency to over-use the extreme response options), and socially desirable responding [58] (some make a distinction between acquiescent and dis-acquiescent, and the extreme and middle response styles [59, 60, 61]. Here acquiescent/dis-acquiescent and extreme/middle-range response styles are considered as two aspects of the similar end-effect).

Response style effects can be especially problematic when different populations are compared and some populations are more susceptible to response styles than others [12]. For example, if a result was obtained showing a difference in lighting preference between women and men, and women were more inclined than men to agree with any assertion the question is making, the result could be just an artefact due to the question format rather than an actual difference of preference between the sexes. The same problem may arise when a study includes comparing the preferences of ethnically different populations.

The theoretical explanations for using response styles have been of either the dispositional (cultural or psychological factors) or situational (the measurement event and its context) variety [59]. Baumgartner et al. assert that stylistic responding is best understood as an interaction of personal dispositions and situational factors, that is, people differ in their inherent tendency to engage in stylistic responding, but this tendency may be encouraged or discouraged by the situation [59]. Response styles are considered to be fairly stable, both over single questionnaire administration, and over longer period [61]. Furthermore, response styles may be greatest when vague, ambiguous or difficult to answer items are involved [48]. The mode of data collection may also influence the severity of response style effects [60].

In social and a political sciences context many studies have shown cultural differences in tendency to acquiesce (e.g. [12, 45, 58, 62]), for extreme response styles [58, 63], and to respond in socially desirable way. These effects persist even when controlling for variables such as the effects of education, the ethnicity of the interviewers [12]. The culture difference does not need to be as wide as for instance, that between the U.S. and Japanese cultures. Method bias can play a significant role in cross-cultural research, even with countries that have a fairly similar standing on major dimensions, like economic affluence, education, and political organization. For example, significant differences in acquiescence have been found in marketing research data e.g. between French and Italian respondents [58].

Response styles cause a measurement error, which leads to differences between a respondent's true opinion and the reported score. This error is generally non-random, i.e. response styles not only contribute to observed score variance, but also may have systematic effects on scale scores and produce correlated errors of measurement [59]. The effect size of differences caused by response styles can be almost as large as that of the reported effect of what was being investigated [20, 58]. Response style effects are included in the common method variance, which is the variance on the data arising from the method used. It should be noted that the common method variance includes other factors than the response style. Cote and Buckley examined the amount of common method variance present in 70 validation studies in the social sciences and estimated that, on average, measures contain 41.7 % trait variance, 26.3 % method variance, and 32 % random error [64]. In other words, less than half of the observed variance is due to the trait being measured. They also found that the amount of variance attributable to method biases varied considerably by discipline; on average, method variance was lowest in marketing research (15.8 %) and highest in education research (30.5 %). Williams et al. investigated the applied psychology literature and found that around 25 % of the variance in these studies was caused by the methods [65].

Response style biases can be controlled through having a sufficiently large sample, careful design of the experiment procedure, and statistical methods. The simplest methods involve just calculating the fraction of the responses that agree and disagree, or the net score of the two, or the proportion of the responses that use the extreme values of the scale [63], but plenty of more complicated procedures have been used (e.g. [59, 60, 62, 66, 67, 68, 69]).

It is not quite clear what would be the best course of action once response style effects are detected. Some researchers assume that tendency for response styles is a personality trait, and the effects can be eliminated by identifying these subjects and removing them from the sample [12, 24]. As Krosnick points out, however, responses are influenced by circumstantial factors as well, not just by personality, and it is impossible to say what the responses would have been [24]. Deleting cases *post hoc* is always something that needs to be done with great care, as it may distort the results and opens a way to manipulate the data to fit better the expectations of the researcher. As response styles are influenced by socio-economical and cultural factors and, sex of the respondents, deleting these cases may distort the results [67]. Such procedure can also only deal with outlier individuals, but does not help if response styles differs between whole sub-populations of the subjects.

4.1 Acquiescence/Dis-acquiescence

The questions in a a questionnaire are often formed so that the subject has to make a “agree–disagree”, “yes–no” or “true–false” judgement [24]. Acquiescence refers to the phenomena that some subjects tend to agree with whatever assertion is made in the question, regardless of the content or the wording of the question [12, 24, 45, 63, 67]. This can create a bias in the resulting data, especially if the questions are phrased so that every “agree” statement lean to the same conclusion. It may also heighten the correlations among unrelated items that are worded in the same direction, and lessen the correlations between items that are conceptually related but worded in opposite directions [20, 59, 67], producing similarities or differences which are due to the method instead of difference of preference [59, 67]. Acquiescence can also create conflicting data, where the subjects agree to mutually contradictory statements. Based on analysing the results of several studies, it has been estimated that the effects of acquiescence is around 10 % in agree–disagree forced choice situation [24], i.e. roughly 10 % more of the subjects agree on an assertion than disagree with the oppositely phrased assertion.

In some cases, e.g. personal animosity towards the researcher or older experienced subject resenting the assertions the younger researcher seems to be making, could lead to a reverse acquiescence or dis-acquiescence. In other words, the subjects would be inclined to *disagree* with everything the questionnaire asserts.

It should be remembered that acquiescence refers agreeing automatically without considering the question. Tendency for positive responses is not an automatic indicator of acquiescence; the questionnaire may have been constructed so that disagreeing responses are less likely. For example, if lighting with known good characteristics is being evaluated, and all the questions are phrased so that the agreeing statements all signify "good" lighting characteristics, then negative responses would be less likely. Acquiescence is also distinct from preferring generally the responses in one side of the range, although it may be impossible to tell which of these two mechanism is responsible.

Several explanations have been offered to explain acquiescence. Acquiescence may be a form of satisficing, offering an easy way to respond without performing the cognitive work involved [24, 34]. It is easier to agree to an assertion without evaluating it than to make a judgement. Most people typically begin judging by seeking reasons to agree rather than disagree. Thus, a subject may become fatigued before getting to the task of generating reasons to disagree with the assertion [21].

Some subjects may treat the assertion in the question itself as a cue at to what the "correct" answer to the question is [12]. A question that is difficult to answer may also trigger acquiescence as a way to escape the embarrassment of admitting ignorance or non-comprehension [12, 24, 34]. It is also known that when people are guessing at true/false questions, they say true more often than false [21]. Some subjects may feel that disagreeing is more impolite than agreeing, or agree to avoid confrontation [12, 24]. This may be either general politeness, or caused by the social situation. For example, in an *in situ* study the subject may feel that the researcher is a guest, and the subject himself is a representative of his organisation, and that one must not be impolite or discourteous to the guests. Such effects are not necessary conscious, so just asking the subjects to be honest may not be enough to remove them.

Acquiescence may also be a form of social deference, if subjects feel themselves to be of lower status than the experimenter [12, 24, 67, 70]. Such perceived social inequality can be caused by ethnic or class differences, or differences in income or education level. In universities, where much of research work is done, the status as student or faculty may be factor – especially if the subjects are recruited from the students and the researchers are members of the faculty who may also be involved in grading or otherwise evaluating the studies of the subjects.

In social sciences research it has been found that subjects with high tendency for acquiescence differ demographically from other respondents, and it is possible that they differ in other ways as well (e.g., ability to follow instructions or ease of being influenced) [12, 67]. Johnson et al. analysed employee satisfaction data from 19 countries and found that subject with a cultural background that stresses more conformity or modesty are more likely to acquiesce than ones with a strong tradition of individualism, and that that acquiescence is less common within cultures that reject ambiguity and uncertainty [45].

Acquiescence is often controlled by including statements with identical content but worded so that the resulting agree–disagree choices are opposite if the judgement remains the same [47, 59, 67]. For example, the questionnaire might contain both statements "The lighting in the left-hand compartment looks more natural than the lighting in the right-hand" and "The lighting in each compartment looks equally natural." The resulting repetition of questions generally improves the reliability of the data even when acquiescence is not occurring. On the other hand, doing this doubles the number of questions making the questionnaire more cumbersome, and it may be very difficult to construct the opposite statements so that they are exactly identical in content [12, 24].

Tendency to acquiesce cannot be assumed to be constant among individual questions, and it is not clear why individuals who acquiesce should be assigned scores at the middle of the response option range [23]. Negative statements may produce artefacts where a relationship appears in the results between the negatively worded items [20]. Cognitive processing of a statement involving a negation is harder, so the negative statement should be made without using the word "not" [24], and double negatives should be avoided at all costs [11]. Reversed items are also known to lead to various problems, including poor model fit of factor models [47] and lower internal consistency [71]. Also, other mechanisms than acquiescence can lead to similar effects. For example, careless responding where the subjects are inattentive and do not notice that some items are keyed reverse; or confirmation bias, where the question prompts the subjects think in ways that are consistent with the stimulus questions [47]. Therefore, misresponse to reversed items is not an automatic signal of acquiescence.

Another signal of acquiescence is a consistent agreement on several items which are not related in content [47, 59]. Pure acquiescence is caused by the stimulus question, so it does not depend, in principle, on how the questions are grouped [47]. However, acquiescence often leads to response-set behaviour, where positive responses are clustered with other positives and negatives with other negatives [67]. Winkler suggest partial correlation analysis to see if the subjects agree on logically contradictory statements [67]. Other methods include inserting neutral or meaningless questions to see to estimate the extent to which the subjects simply agree to every questions. A procedure favoured in some fields is to average the agreements that a respondent records with a full range of conceptually unrelated items and subtracting this mean from responses to individual items to derive scores free of acquiescent bias [62]. However, scores computed in this way are not statistically independent of one another, which limits the subsequent statistical analysis [62].

The most effective way to avoid acquiescence is to avoid agree–disagree judgements altogether [23, 24], and instead to ask questions in a balanced manner with substantive response alternatives rather than with just agree-disagree choice [12, 24]. Forced choice alternatives have also been suggested [23]. However, few methodology articles goes this far on their recommendations [12]. This would effectively require abandoning Likert-type response formats completely. A more common approach is to try to find out which statements or questions are most likely to arouse acquiescence and avoid them [12].

4.2 Extreme response style/middle response style

Extreme response style (ERS) is the tendency to use mostly the extreme response options of a range [45, 58, 59, 63, 68, 69]. For example, a person might respond always with either “fully agree” or “fully disagree avoiding the “somewhat” options, or only use the extreme points on a numerical scale. The middle response style (MRS), also called response contraction bias, is the converse of the ERS, i.e. the tendency to always respond with the middle or qualified answer. Some claim that the MRS is the more prevalent of the two [8].

It should be noted that an individual responding with mostly the extreme (or middle) values is not necessarily responding dishonestly or satisficing (although he may be). Indeed, the interpretation of ERS is an interesting question; do the extreme values indicate more strongly felt opinions, or do those who tend to respond with extreme values exaggerate their opinions (and do those who never use the extreme options downplay their opinions)? There is no clear answer to this question [63]. If the respondent using ERS truly has stronger opinions, then the extreme values are accurate assessments and not false, unlike distortions caused by acquiescence or socially desirable responding.

Extreme responding may also be one-sided; such subject would have a tendency to always “strongly agree” when he agrees, but only “somewhat disagree”, never using the extreme disagreement end of the scale no matter what. Such effect would most likely be explained by acquiescence or socially desirable responding style effects; expressing strong disagreement can be more difficult if it is seen as impolite, while strong agreement may be a signal of earnestness or having common ground. However, it has been found that the tendency to use one extreme end of the scale correlates positively with the tendency to use the other [63].

Extreme response style creates a bias in the data, which can affect the location of the central tendency (means or medians). If two populations are compared so that one of the populations has a greater tendency for extreme response style (due to e.g. cultural difference), the results can show false differences between the populations. Extreme response style also leads to increased variances, which may reduce the power of subsequent statistical analysis. The extreme case is that ERS reduces the many-point response range to a binary choice between the two extreme alternatives. If the opposite happens and the subjects avoid the extreme values, the observed effects will appear smaller, which will reduce the statistical power. Baumgarter and Steenkamp suggest that researchers should be particularly concerned about ERS when respondents tend to have relatively strong opinions (in either a positive or negative direction) about an issue [59].

Several reasons for using different amounts extreme response style have been offered. The most commonly studied factors comprise cultural norms, the psychological factors of the subjects, and the measurement event itself (e.g. different response formats can produce different tendency to use ERS) [72]. Most likely the ERS is influenced by interaction of all three.

ERS can be stronger in cultures where earnestness, sincerity, conviction are highly valued, whereas in cultures where modesty and harmony are emphasized the middle values of scales are used more [72]. Extreme response style helps to achieve clarity, precision, and decisiveness, which are valued in masculine and high power distance cultures [45]. Chen, Lee and Stevenson found that Japanese and Taiwanese students used more the middle value of a 7-point Likert response format than American and Canadian students [73], and that the US group used more extreme values than the Japanese, Taiwanese or Canadian students [73]. In an analysis of European market data, van Herk et al. found that the extreme response index is higher for Greek than for Italian and French respondents, and the extreme response index was higher in Italy than in France [58]. Italian respondents consistently tended to have a higher extreme response index than respondents in the other Western European countries, and the extreme response index in Spain was consistently higher than in France, Germany, and the United Kingdom [58].

There are within-nation differences, too: Hispanic American students have been found to use more the extreme points than Caucasian American students [74]. Interestingly, studies using bilingual respondents have indicated that ratings also vary with language of response. For instance, Hispanics have showed more extreme responses when completing questionnaires in Spanish than in English [62]. This may indicate that ERS may be affected by the linguistic style [63], and that “strongly agree” is not equally strong in different linguistic settings.

There are also substantial and consistent individual differences between the tendency to use or avoid the extreme ends of a scale [63]. ERS has been explained intolerance of ambiguity and dogmatism [59]. Sometimes the respondents may think that a “normal” person falls in the middle of the scale, and use the opinions near the middle in order to present themselves as “normal” [21, 59]. Middle response style may also result from the subject not wanting to express his true opinion, thus the mild or neutral options forming a “safe” response [59].

Other researchers have argued that some features of rating scales might connect to respondents’ tendency to choose extreme positions on rating scales. A significant amount of ERS has been reported for a 5-point response range, whereas the amount was negligible when using a 10-point range [72]. Providing labels for the endpoints but not the midpoints of a response scale, may be more problematic for respondents with an intolerance for ambiguity and may encourage them to select extreme response options, which are more clearly labelled [45]. Weijters et al. found that labelling all the response options in a Likert instrument reduced the tendency for extreme responses, but increased net acquiescence when compared with the case where only the end-points were labelled [43]. They also found that including a mid-point (neutral) response option decreased the occurrence of ERS [43].

The simplest way to detect ERS is to calculate the number or proportion of responses to heterogeneous questions with the extreme option is selected as the response [59].

4.3 Socially desirable response style

Socially desirable response style is the tendency of subjects to present themselves in an (overtly) favourable light, regardless of their true feelings about the question they are being asked [20, 57, 58, 75]. This may be either over-reporting of admirable attitudes and behaviours, or under-reporting those that are not socially respected [21]. This can not only cause bias or mask true and create and false correlations in the responses [20], but in the worst case the subject is responding in a way that has nothing to do with his real opinion on the question. One could also argue that the acquiescence or extreme response style biases could at least in part be due to social desirability effects [58].

Socially desirable biases are often explained as intentional misrepresentation by respondents, but they may be at least partly caused by simple mistakes in recall, particularly if time has passed between since the even to be recalled [21, 24]. Often, these biases are unconscious, rather than a deliberate deception [13].

The subjects are usually concerned how well they are performing in the test, which can be regarded as an aspect of socially desirable behaviour. The subject's idea of good performance is often different from the researchers – The researchers wishes that the subjects respond with honest, unbiased evaluations, whereas the subjects often assume that there is a correct way to answer and thus seek cues for what they "ought" to answer. The subjects also have a desire to appear consistent and rational in their responses and might search for similarities in the questions asked of them, thereby producing relationships that would not otherwise exist at the same level in real-life settings [20].

Respondents may form expectations about what is being measured, and respond to individual items based on their overall position concerning the focal issue, rather than specific item content. This may happen because the instructions may tell respondents what the researcher is interested in, or indicate it with a heading or label in the questionnaire. A respondent may also infer what is being measured from the questionnaire; related items are often grouped together. [47]

It would be easy to dismiss the socially desirable response effects from lighting preference research. One can easily argue, that the subjects may want to answer in socially desirable way about a question on, say, how healthy dietary habits they have, but they cannot make themselves look better by answering whether plastic fruits look more natural in the left-hand or right-hand compartment. However, questions often include items about the current mental state of the respondents (e.g. mood, tiredness or alertness), and answers to these questions can very easily be influenced by one end of the scale being more socially desirable than the other. Environmental consciousness may also influence the data via promoting socially desirable answers, particularly if the respondents know what they are comparing (and it can be impossible to disguise the yellow-green light of an sodium lamp in street lighting preference study, for example).

4.4 Conclusions

This and the preceding chapters have discussed some –but is by no means a complete list– of potential bias sources. A large part of the bias-related research is from fields other than lighting research, and it is a fair question to ask if these biases are relevant in lighting acceptance and preference studies. Much of lighting acceptance and preference research is done by comparing the data from two or more lighting settings, i.e. what is interesting is the difference between the ratings, not the values themselves (which usually mean very little by themselves anyway [4]). It could thus be argued that many biases described above create an approximately constant term to the results, which cancels out when the results are analysed by comparing them. This view must be regarded as too simplistic. A number of observations can be made:

First, it is intuitively clear that anything that causes the responses to deviate from the true opinion will make the data less valid and reliable.

Second, effects such as acquiescent or socially desirable response styles can lead to conflicting data, making it more difficult to make definite conclusions.

Third, response style effects can lead to a situation where the number of scale options are in effect reduced. Acquiescence may lead to situation where the stronger disagree options are never used, and extreme/middle response style both lead to situation where a part of response options are effectively omitted. This may effect the resulting data characteristics negatively or hide the more subtle differences.

Fourth, biases generally do effect the statistical analysis of the data. They can contribute to the variance, lower the statistical power of the study, and can create false correlations or hide true ones.

Fifth, the possibility of different response style effects must be kept in mind if the research results are showing a difference between demographically different populations. There may of course be genuine differences in preference, but in these cases at least a rudimentary analysis should be performed to see to what extent response styles can explain the observed differences.

Sixth, the researcher needs to keep the possibility for these effects in mind when comparing his results with those of others, or comparing several studies from literature.

Because of these factors, it is obvious that avoiding or lessening response style effects can be expected to improve the characteristics of data. The research instrument, setting and protocol can be designed to avoid these effects, which will improve the quality of the research. Usually this does not even require significant extra work, just paying attention to the experimental design. Acquiescence effects can be lessened by avoiding agree–disagree decisions and offering substantive response options instead. The motivation of the subjects should also be paid attention to, as well as stressing that there is no “correct” way to answer.

5. Statistical analysis of questionnaire data – a review of a number of controversies

Statistical analysis techniques seem now days to be almost a requirement of doing science. Statistics is a field advancing rapidly, but most non-statisticians stick to the few basic methods. Statistics is taught, at least to engineers, as a branch of applied mathematics with applications of its own in various engineering disciplines such as quality control. The introductory statistics courses -often the only mandatory ones- tend to present the subject as a simple, uniform and non-controversial subject. Several writers have noticed, that textbooks and guidelines often contain errors or oversimplifications making statistics appear as rigid rules to be followed [76, 77, 78].

Statistics often involve a distinct hard-soft dichotomy. The methods involve complex calculations involving intimidating-looking equations, which are based on properties of distributions that are somewhat arcane (at least to non-statistician). Statistical methods will give as results numerical data with many decimals, which creates the appearance of great precision and exactness. But, these methods involve also very “soft” components; they are robust if the actual data conforms “reasonably” with the assumptions the methods make, or the sample size is “large enough.” The guidelines of determining what is statistically significant or when an effect is large or small, are based on convention, rather than on some hard objective properties of the methods involved. Instruction texts include often phrases such as “..if the distribution is reasonably normal (for “reasonable, consult your local statistician)...”, but the local statistician cannot often give very clear answers. This is not surprising, for if there were such simple rules, they would have been printed on the textbook in the first place rather than a suggestion to consult the statistician.

Statistical practices involve, however, a number of deep-set and enduring controversies. Texts intended for practical guides or rules of thumb habitually omit to mention that there are controversies. This chapter will review three persistent controversies of statistics: null-hypothesis significance testing, using parametric methods to analyse ordinal data, using unmodified tests for multiple comparisons, and using parametric methods to analyse ordinal data.

5.1 Null hypothesis significance testing – criticism and counter-criticism

Statistical null hypothesis significance testing (NHST) is very widely used to evaluate the reliability of the results produced by a study. It is used in practically all the scientific fields. For example, it is estimated that 94 % of empirical papers published by the American Psychological Association (APA) journals used significance tests [79].

There has been a lively controversy of using statistical hypothesis testing ranging for decades with vocal arguments that it should be banned, abandoned completely [80] and replaced with something else. Yet there is little evidence that the discussion or criticisms have reduced its popularity among researchers [76, 81].

The discussion has been at times quite passionate, as illustrated by some of the titles of the papers taking place in the controversy, ranging from polemic (title, that is, the article itself is quite agreeable on style) “The ongoing tyranny of statistical significance testing in biomedical research [82]” to fawning “In praise of the null hypothesis statistical test [83]” to humorous “The earth is round ($p < .05$) [84]”.

Unfortunately, a researcher is not completely free as to whether to use NHST or not, as editors and reviewers often demand it. Therefore, just ignoring it and always using something else is not likely to be a fruitful approach.

5.1.1 Null hypothesis significance testing

Null-hypothesis significance testing, as practised today, is an amalgamation of work of Fisher, and of Neyman and Pearson [85, 86, 87]. They had different approaches; Fisher was a scientist and treated NHTS as incremental, driven by replication, improving with each NHST decision, and potentially self-correcting [81, 85], whereas Neyman and Pearson thought more in terms of decision-making in practical applications (e.g. quality control), where each decision is important [85]. Fisher was not interested in alternative hypotheses, which were the focus of Neyman and Pearson, who in turn were not thinking in terms of proof by contradiction [85, 86]. The current practice combines some of the features of both approaches, with most texts implying that NHST is bases on a unified well-tested theory [78, 79, 81, 82, 86, 88, 89, 90, 91]. The names and history are often omitted, and the major ideas are presented anonymously, as if they were given truths [78, 88]. The very purpose of NHST is often left unstated; is it performed to learn something about the subject of the study, or is it something to evaluate the experiment by?

Classical statistical null hypothesis testing includes four steps

1. State the research hypothesis and the null hypothesis

2. Gather data

3. Conduct a statistical test which should result in a p-value

4. Make inferences on the null hypothesis on based on the p-value

Each of these steps are important. The first step, to explicitly state the research hypothesis and the null hypothesis is often omitted in practice, and the published hypotheses are those that arise after the results are analysed.

The third step, a statistical test, can occur via many techniques. What is common to all is to calculate a test statistics, which is compared with the assumption that the samples were all drawn independently from the same underlying population.

The index p (sometimes written with a capital P) is defined as *the probability of obtaining the observed result, or an more extreme result, if the null hypothesis were true,* or in mathematical form

$$p = \Pr(D \mid H_0), \tag{5.1}$$

where

D is the value (or greater) of the test statistics, and

H_0 is the case where null-hypothesis is true

The value of p depends on two factors: size of effect, i.e. how different the

		Actual state of the null hypothesis	
		True	False
Decision	Rejected	Type I error (α)	Correct decision $(1-\beta)$
	Not rejected	Correct decision $(1-\alpha)$	Type II error (β)

Figure 5.1. The possible outcomes of NHTS.

observed result is from the null hypothesis and how great the variability is in the data, and the sample size.

The last step, making inferences from the p-value, is a subtle one. Although the numerical value is calculated with great precision, the inference made from it is an inductive one: if the p-value indicates that the probability of obtaining this result is small assuming that the null-hypothesis was true, it is concluded that the assumption of null hypothesis is true is likely to be wrong. But there is no precise figure for telling how likely it is wrong, and the question of how unlikely the result is under the alternative hypothesis is not addressed at all. The calculated p-value gives no information on how unlikely the result was if the test hypothesis were true. The p-value in itself is not the probability that the effects arose by chance, or that there is no effect, or of the conclusion being wrong.

The null-hypothesis can be either true or false, and the experimenter can decide either to reject or not reject it. It must be stressed that "not reject" does not mean "accept" – most statistical procedures cannot be used to show that the null hypothesis is true. This generates four possibilities as shown in the figure 5.1.: Either a correct decision is made, or an error of Type I or Type II is committed. NHST is intended to guard about Type I errors, i.e. false positives, when the null hypothesis is rejected when it is in fact true. This is done by setting the criterion α, typically 0.01 or 0.05, and rejecting the null hypothesis only if the observed p is at most α. The other type of error, Type II, occurs when there is a real effect, but the decision is made to not reject the null hypothesis. The typical case of the Type II error is that there is a real effect, but the experiment fails to detect it. The parameter β is the probability of not rejecting the null hypothesis when it is, in fact, false. Statistical power is the probability of correctly rejecting the null hypothesis, or $(1-\beta)$. The value of β, like that of α is set by convention, the most commonly recommended combination being $\alpha = 0.05; \beta = 0.2$ (and thereof power 0.8) [92].

It should be noted that the rates of Type I and Type II errors are interrelated; if it is wished to keep guard more stringently against Type I

		Statistical significance test	
		Significant	Not significant
Importance of the result	Important	Elated	Very sad
	Not important	Annoyed	Happy

Figure 5.2. States of mind of an hypothesis tester, according to Nester [93].

errors by setting the value of α smaller, the criterion for rejecting the null hypothesis is made more demanding. This will increase the Type II error rate if there is an effect, as the null hypothesis are now not rejected in some cases where they would have been be rejected when the α was higher.

There can be two reasons for the NHST returning "not significant": either the effect is too small, or the statistical power of the test is too small. In the former case, the interpretation depends on whether the detected effect size is large enough to have practical or theoretical importance. If yes, the researcher is very sad, because he now needs to repeat his experiment with a larger sample size (simply collecting more data and combining it with the data already on hand is considered poor practice); if not, then the researcher is probably happy that he did not waste more time, effort and money in investigating something that turned out to be of no interest. Similarly, a statistical significant result can make the researcher either elated or annoyed, depending on whether the detected effect has practical or theoretical importance. A statistically significant effect of practically important size is a potential "Eureka" moment: something important has been found or demonstrated. But if the detected effect size is too small to have practical or theoretical meaning, then the researcher is annoyed having an irrelevant finding (Figure 5.2.).

As Nester pointed out, if the researcher is testing some pet theory then an additional layer of complexity arises. A statistically significant result is either good or bad from the point of view of the experimenter, depending on how well the result fits his theory, or how well he can explain a result that does not. Similarly, a statistically non-significant result can be discussed as good when it fits the expectations of the researcher (when the non-significance is usually interpreted incorrectly as no effect), or bad if the effect is what was expected, but the sample too small. [93]

5.1.2 Arguments against using the null hypothesis significance testing, and their counter-arguments

Although many of the points raised by those who argue against the use of NHST have remained unchallenged [78, 94], the procedure has become more or less the standard way to provide criteria for separating signal from noise in the majority of published research [95]. This may also explain why the defenders of NHST generally do not present strong arguments for adopting it. Rather, they mostly address the arguments of those who argue for limiting its use.

Therefore, the following discussion about the arguments is presented by presenting the arguments against using NHST, with their counterarguments. This has the effect of making the use of NHST to appear as the default position. This position is not intended to imply any inherent superiority of NHST, but merely the fact that it is, in practice, the most widely used method. Nor is the appearance of an counterargument after the argument to be taken to suggest that the counterargument is conclusive; the purpose is only to present the arguments and counterarguments.

Null hypothesis testing does not provide any information on the phenomenon being investigated

Test of significance does not help to estimate the value of a parameter, nor does it provide insight to the phenomenon or attribute being investigated [76, 90, 94, 96, 97]. If there is a real difference, no matter how small, then the null hypothesis is false and the result can be made statistically significant by using a large enough sample [85, 87, 94, 96, 98]. Thus, the decision of whether or not the null hypothesis is retained is in practice reduced to

1. The null hypothesis was rejected because enough samples were collected.

2. The null hypothesis was not rejected, because not enough samples were collected. [93]

Thus, statistical significance can be regarded largely as an indication of the sufficiency of the sample size –and a rather complicated and non-intuitive way to indicate this. Since the statistical significance or lack of it depend so much on the sample size, they give no information on the phenomenon being investigated. In fact, if conclusions were drawn from statistical significance and lack of it, it would be misleading because these conclusions depend on the decisions of the researcher made when he decided how large sample he is going to use [76].

The counterargument to this is to adopt the position that the aim of NHST is not to learn about the phenomenon, but to evaluate whether the study justifies making a statement about the phenomenon or attribute. Thus, to properly report the results, one need to tell what they are (direction, magnitude, effect size, etc.) and show with statistical methods that there is sufficient reason to accept them.

The term statistically significant is highly misleading and has nothing to do with the noteworthiness of the results

Practically every text arguing against NHST objects the use of word "significance." The word "significant" is misleading as it refers only to the world of statistics [76, 81, 82, 85]; statistically significant is not necessarily be noteworthy in practice. Despite this, statistical significance is frequently seen as practical significance when it in fact refers only to the rejection of the null hypothesis. Therefore, irrelevant results which were statistically significant are presented as discoveries, and interesting results with notable effect sizes are ignored or not published if they are not statistically significant [82, 93, 94].

In literature the word "statistical" is often omitted from the term "statistically significant", which increases the confusion [94]. For example, Dawes studied if the data characteristics change when different number of response options are used [26]. He compared Likert-type data with either five, seven or ten response categories. Dawes stated his findings in his original paper (emphasis added):

> "To test if the overall mean scores of the eight items were statistically significantly different...a one-way ANOVA was run...The results was *statistically significant.* Based on this result, it seems that a 10-point scale will produce *slightly* lower scores compared to scores generated from 5-point or 7-point formats..."

Now, compare this with the way this is referred to in another paper [9] (emphasis added):

> "...and this revealed *significant* effects on the mean rating..."

and later in the same paper:

> "...suggested *significant* effects on the mean rating, but not on dispersion."

Very careful reading is needed to understand that the latter paper refer only to the statistical and not practical significance, especially since it makes no reference to the fact that Dawes considered the difference to be slight.

Neither Dawes nor the paper referring to him considered the effect size, but it can be easily estimated from Dawes' results. Since the differences between the re-scaled means was 0.3, and standard deviations 2.0 or 1.9, Cohen's d can be calculated to be 0.15–0.16. In other words, the effect size was small and Dawes' characterisation of the results as "slightly lower" is accurate.

Null hypothesis testing involves a logical fallacy

The inductive nature of NHST makes it possible to argue that it entails a logical fallacy of affirming the consequence, i.e. having the rule "If A then B", observing B and then concluding A [76, 84]. This is fallacious, because there may be other causes of B than A.

NHST makes use of similar reasoning; it calculates the probability of data given the null hypothesis, which is used to make inference of the null hypothesis given the data. As Cohen put it:

> "It [NHST] does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" [84]

To infer the confirmation of the alternative hypothesis from the rejection of the null hypothesis has been called similarly fallacious.

NHST has been likened to the *modus tollens* form of argument: "If A, then B; observed not B, therefore, not A." This is a valid form of logic if the premises are categorical, i.e. A would always result into B, but invalid when the premises are only probabilistic [76].

This argument can be criticised by that it applies the concept of fallacy beyond its domain. The statements involved in the null hypothesis test are probabilistic, and the reasoning involved is inductive, thereby their form includes the possibility that the conclusion may be wrong. The situation is outside the bounds where the rules of formal logic apply. Science is fundamentally an inductive process; the usual way to test a scientific theory is to make predictions based on it, and to see whether these predictions prove to be true [76]. Thus, NHST is in line with the generally accepted scientific thinking.

Null hypothesis testing tells nothing about the probabilities of the hypotheses being true

Since the p-value is calculated assuming the null hypothesis is true, it cannot be a measure of probability of it being true [81, 94, 99]. Therefore the test of significance does not allow making any statement about the hypotheses in terms of mathematical probability [88, 95].

The calculation of the posterior probability of the null hypothesis being true given the data requires the Bayesian approach instead of NHST. It can be calculated with the Bayesian

$$\Pr(H_0 \mid D) = \frac{\Pr(D \mid H_0) \cdot \Pr(H_0)}{\Pr(D \mid H_0)\Pr(H_0) + \Pr(D \mid H_A)\Pr(H_A)}, \tag{5.2}$$

where

D is the value (or greater) of the test statistics,

H_0 is the case where null-hypothesis is true and,

H_A is the case where the alternative hypothesis is true.

Unfortunately, the results depend on the prior probabilities $\Pr(H_0)$ and $\Pr(H_A)$ (usually treated as it was $1 - \Pr(H_0)$), which are often unknown and sometimes unknowable [95] (or trivially zero and one if the argument that the null hypothesis is usually known to be false *a priori* is accepted, see below). Likewise, $\Pr(D \mid H_A)$ is usually unknown, and because often the H_A is not even specified in any exact way, there may be little basis for making any assumptions about it.

Rejecting the null does not provide logical support for the alternative [85]. Yet it is a widespread belief that the level of significance is the probability that it is correct decision to reject the null hypothesis [76, 78, 81, 84] and adopt the alternative hypothesis by default.

These arguments can be countered by again adopting the stance that the point of null hypothesis testing is not finding if a claim is true but if the evidence is sufficient to warrant claiming otherwise [99]. In this view, the main purpose of NHST is to evaluate the study or experiment, not the results.

Furthermore, Nickerson argued from calculated examples that if it can be assumed that $\Pr(H_A) \geq \Pr(H_0)$ and $\Pr(D \mid H_A) \gg \Pr(D \mid H_0)$, then a small value of $\Pr(D \mid H_0)$ i.e. the p-value, can be taken as a proxy for a relatively small value of $\Pr(H_0 \mid D)$. These assumptions are not unreasonable for much of research work, but is not hard to think of cases in which the probability of a given result would be very small under either the null or the alternative hypothesis. [76]

Null hypothesis is known to be false even before any experiments are made.

An often-stated argument against using NHST is that testing for the null hypothesis is pointless because it is known to be false before any data is gathered [76, 86, 90, 96, 98, 100]. Experimental set-ups are made so that there are differences (it must be kept in mind, however, that the actual difference between the conditions is not necessarily what the research has in mind). If the statistical null hypothesis is almost never true and this is known beforehand, testing to see if it is true seems to serve no purpose.

NHST is supposedly done to avoid committing Type I error (rejection of null hypothesis when it is in fact true), but this error is impossible to commit if the null hypothesis is known to be false . This renders the whole operation again pointless. Rather, focus should be placed on avoiding the Type II error (not rejecting the null hypothesis when it is false), since it is the only type of error that can be made when the null hypothesis is known to be false. NHST does not, however, address Type II errors. [80]

It should be noted that this argument applies only to NHST as practised in scientific studies where the data is formed from a small sample of the general population; in other applications such as quality control or clinical diagnosis the null hypothesis may be true and in fact represent the majority of cases.

This raises the question of what the purpose of NHST is? Is it used to determine if the null hypothesis is true, or is it used to judge not if the null hypothesis *is* true, but whether, in the view of the data obtained, it *could* be true. In other words, does it evaluate the phenomenon or the experiment. If former, the question asked is “is there an effect”, is the latter “do the results of the experiment provide sufficient evidence for claiming an effect”. A not rejected null hypothesis, when the effect size is practically meaningful, means usually that the sample size is too small.

Also, the statistical null hypothesis is not that there is no difference, but that the samples were drawn from the same population. Even means calculated from samples drawn from the same population will have some variation, so that there is a difference does not automatically indicate that the samples were drawn from different populations. To conclude that the samples represent different populations requires that the magnitude of the difference between the means us great enough relative to the standard error [76]. If the samples really were drawn from the same population, the likelihood that the null hypothesis will be rejected does not become a certainty as sample size increases [76, 83, 98].

The null hypothesis does not need to be the unrealistic requirement that there is exactly zero difference, but the exactly zero difference can serve as a proxy for a null hypothesis of a very small difference. Thus, the practical null hypothesis can be that the effect is close enough to zero that it is not interesting in practice. Testing for a point (exactly zero) null is easier than for a null hypothesis of a small range close to zero, and if the sample size is not very large, leads to approximately same results. [76]

Null hypothesis testing is arbitrary

The usual convention is to use fixed level for the criterion for significance or α, often 0.05, to make a distinction between a statistically significant ($p \leq \alpha$) and non-significant ($p > \alpha$) finding. This criterion is not founded in any theoretical property of statistics or the subject under study, but is simply a convention [90, 98]. The real difference between result for which $p = 0.045$ and another for which $p = 0.055$ is very small, so it is misleading to think one as "significant" and another as "insignificant." [76, 80, 99]. On the other hand, if the fixed $\alpha = 0.05$, then the not statistically significant result with $p = 0.06$ is treated in the same way as one with $p = 0.9$ even if the two are very different [99].

The p-value is often misinterpreted as a measure of magnitude [76, 86], so it is easy to find phrases like "almost significant" or "highly significant" in the literature.

Null hypothesis cannot be proven; failing to reject the null hypothesis does not provide support for the it being true

The most often used forms of statistical testing cannot prove the null hypothesis, only disprove it [77, 98]. Failing to reject the null hypothesis can mean either that there was no effect (null hypothesis is true) or that the sample size was inadequate. Thus, if the null hypothesis is not rejected, the only valid inference is that nothing can be said. Care must be taken to not implicitly accept null hypotheses, especially when the sample size is small [98]. Yet several authors point out that the literature abounds cases where the result "no statistical difference" has been used as proof for similarity or lack of effect [76, 77, 80, 81, 82, 98], or marginally significant results are discussed as trends and then treated as facts [77]. Many researchers assume that the true difference is the observed difference, or zero, depending on the outcome of the significance test and the prior expectations [93].

For example, Atli and Fotios [9] conducted experiments to investigate if the number of response categories effect the data obtained in indoor environment studies. Their subjects rated their environment on two different days. Atli and Fotios wanted to combine the data sets from the different days for their analysis. Therefore they conducted Mann-Whitney (MW) and Kolmogorov-Smirnov (KS) tests to see if there was a difference between the ratings done in different days. They then concluded that the data from the two days can be safely combined into one set, because (emphasis added)

> "The Mann-Whitney test suggest *differences* between the evaluation sessions in only two of the 16 cases...[and] The Kolmogorov-Smirnov test does not suggest any *differences to be significant*....it was concluded that *similar response were gained* on both evaluation sessions...[9]"

It is to the merit of the authors to have considered the possibility of an effect arising from the experiments concluded on different days. Unfortunately, their analysis has several problems. The most fundamental problem is that the statistical tests employed in this case can only be used can return only a result of "statistically significant" or "not statistically significant." While the former can be used to conclude that there is a difference, the latter signals only that no conclusions can be drawn. Thus, the conclusion that the response sets were similar is not one that can be made on basis on these tests.

Another problem is that statistical significance depends on the sample size, and the different/similar -conclusion drawn solely from the lack statistical significance would have been reversed if a great enough sample size had been employed. The sample size used is rather small, so the conclusion could well arise from the low statistical power achieved.

They also conclude that there was no differences, even when some of the tests conducted indicated statistically significant differences. The authors justify their conclusion by observing that the Kolmogorov-Smirnov test (which indicated no statistical differences) tends to have better power than the Mann-Whitney when the sample size is small [9]. However, the allegedly superior power of KS test is not a reason to dismiss the results of the MW-test; statistical power is the probability to reject a false null hypothesis (i.e. good power guards against a false negative), and thus it cannot be used to argue that a statistically significant result is a false positive (which is controlled by the NHST). Perhaps the authors reasoned that if only one of the tests shows statistically significant differences, it should have been the more powerful KS. This would be valid reasoning if KS was known to be *always* more powerful than MW, but since it only *tends* to be, the possibility that this was one of the cases where the MW happened to more powerful cannot be discounted. However, when many comparisons are made, and just a few show statistically significant differences, it is prudent to be sceptical (see below for the discussion of multiple comparisons).

A more fruitful approach would have been to estimate the effect sizes from the daily differences, and comparing those with the size of effect resulting from what they were investigating. It is easy to calculate Cohen's d's from the data Atli and Fotios present in their table 1 [9] (Since Atli and Fotios used MW and KS, it would, of course, be more consistent to use the effect size appropriate for these methods. However, the way the data is presented in provides the data in the form to make it easy to calculate Cohen's ds).

When this is done, the result suggests that there may have been a systematic effect so that the ratings were higher on the second day. The Cohen's d's thus calculated ranged from -0.07 to 0.94 (negative score signifying that the ratings were higher on the first day and positive that they were higher on the second), with the overall mean 0.34, indicating a small-medium effect size between the days. There were few (2 out of 16 comparisons) individual cases where the first day ratings were slightly higher, but the mean effect was to the same direction (day 2 higher) in every case when the mean of effect sizes was calculated either question-wise or response range -wise.

As there were two cases out of sixteen where the authors reported MW test indicating significance, and a clear pattern emerges from the effect sizes, it would not be unreasonable to suspect that a possible day-effect may have occurred.

Null hypothesis testing is confusing and widely misunderstood

NHST is poorly understood by many of the researchers who use it. Empirical results show that many researchers do not understand the logic and limitations of the statistical tests, and hold many misconceptions [79, 81, 91, 97], and draw on the basis of test results conclusions that the data do not justify [76, 79, 81]. The average student cannot describe the underlying idea of null hypothesis testing; What the students learn is only the mechanical process of calculating a significance test [78]. It has been pointed out that the misconceptions are not only held by novice or mediocre researchers, but by some of most distinguished authors, researchers, editors and reviewers in the field of psychometrics [77, 101].

For example, Haller and Krauss conducted a survey experiment on 30 statistics teachers, 39 psychology teachers (not teaching statistics), and 44 psychology students from the psychology departments at six German universities [78]. The survey contained six statements regarding NHST inference, with an agree or disagree choice. The statements were all false, containing some of the most common misconceptions of NHST. Each statistics teacher taught null hypothesis testing, and each student had successfully passed one or more statistics courses in which it was taught. The result was that 100 % of psychology students, 90 % of the psychology teachers and 80 % of the statistics teachers agreed to at least one of the six stated misconceptions of hypothesis testing. Even if this result seems extreme, it is in fact somewhat better than that of Oakes, who conducted the same experiment in the US [78].

The most obvious counterargument is to point out that widely held misconceptions do not invalidate the NHST as a method. Instead, care should be taken that it is used correctly[76, 77].

Emphasis on null hypothesis significance testing has led to poor, even harmful practice

Since the result of a null hypothesis significance testing is dichotomous decision between statistically significant and non-significant, it steers analysis towards similar yes-no treatment instead of magnitude estimation or model-creation [76, 94]. This is not only poor, but harmful; since statistical significance depends greatly on the sample size, the decision and thus findings becomes arbitrary. Since the alternative hypothesis needs not to be clearly specified in NHST, there is no incentive to do so, which hinders the researcher from developing theories or models [76].

The convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance leads to a high rate of non-replication (lack of confirmation) of research discoveries [101].

Null hypothesis testing contributes to publication bias

Publication bias arises whenever the probability that a study is published depends on the statistical significance of its results [102]. If the author does not obtain significant results, the likelihood of being published is severely diminished [76, 86, 94, 102], either because of editor bias or because the researchers do not submit non-significant findings at the first place. Thus there may be thousands of studies with meaningful effect sizes that have been rejected for publication or never submitted for publication [76, 86, 102]. Instead of all research effort contributing to the body of research knowledge, only the studies that are lucky enough to reach statistical significance via large sample size, or via chance, ever reach the research community [86].

This has led to many adverse effects. The emphasis on the type I errors may have contributed to the lack of power in published studies [86]. Scientific inquiry can be retarded because many worthwhile research projects cannot be conducted, since the sample sizes required to achieve adequate power may be impractically large [86]. Publication bias also distorts the results of combining the results of many studies, because the studies combined from the literature is not a representative sample [102]. Publication bias causes the evidence against the null hypothesis appear stronger than it actually is. If only or mostly statistically significant results are published, the published results are likely to involve much more false rejection of null hypotheses than the p-values would indicate [76, 96].

The publication bias also leads to the reported effect sizes being inflated from those in the actual population. Since sample effect sizes need often to be larger than what is in the population to be significant, the studies which overestimate the effect size are more likely to be published [84].

Null-hypothesis testing has become a mindless ritual which is mechanistically followed

A very common observation is that NHST is performed mechanistically as a mindless ritual [84, 88, 93]. Statistical methods are taught as a set of rules to be followed, at the cost of common sense and independent thinking [89]. Data analysis is, however, art, not algorithm [100].

The interesting question is why NHST has become a mindless ritual. The most commonly cited reasons for popularity of NHST are:

1. It appears to be simple, objective and exact [86, 93, 103], and are seen by some as a way of standardisation in the interest of objectivity [76]. NHST also provides a common point of reference between many statistical methods [76, 99]. Testing is seen as a system of objective, automated decision-making [103].

2. It is used as a stamp for approval [96]; complicated mathematical procedures lend an air of scientific objectivity to conclusions [98].

3. It is readily available in many commercial statistics software packages [86, 93, 103].

4. Everyone else seems to use NHST [93] to the point that it has become a tradition[98]. Researchers from other fields than statistics are likely to simply follow procedures learned from persons assumed to know more of statistics [76]. Most researchers are not really interested in statistical thinking, but in their own field and how to get their papers published [88].

5. Students, statisticians and scientists are taught to use NHST [93], and are ignorant of the alternatives [98], or not confident enough to use them.

6. Some journal editors and thesis supervisors demand them [86, 88, 93, 96]. The smart graduate student does not want problems with his instructor, the smart post-doc needs to publish quickly to secure a tenure, and the smart professor needs to see that his students and post-docs graduate and publish quickly [88].

If the fundamental reasons for the mechanistic use of NHST are not addressed, it is all too likely that whatever replaced NHST would be applied just as ritually – and it is questionable whether the benefits reaped from changing the liturgy justify the inconvenience of doing so.

5.1.3 Alternatives to hypothesis testing

It has been proposed that researchers should move the emphasis from NHST to building statistical and scientific models [76, 85]. This facilitates better the understanding of the phenomenon being analysed, and estimating the parameters of interest.

Effect sizes measures and measures of association have been proposed as alternative or complement to NHST [76, 80, 86, 87]. The effect sizes of several studies can be evaluated together with meta-analysis techniques [80].

However, reporting only effect size is not sufficient to separate the true findings from the chance ones [104], and therefore it is recommended that both effect sizes and standard errors should be reported [76]. It is not always clear which of the many possible effect size indices is most appropriate in a given situation, and an effect size that is large or a small by statistical convention does not necessarily indicate that the observed effect is notable or not important in practice [76]. The detected effect sizes are only estimates, with uncertainties of their own. Therefore, it has been suggested that the confidence intervals of the effect sizes should be also presented [105].

Bayesian statistics offer an alternative approach [76, 81, 98, 99]. One Bayesian approach is to compare the posterior odd ratios [76]:

$$\frac{\Pr(H_0)_{post}}{\Pr(H_A)_{post}} = \frac{\Pr(D|H_0)}{\Pr(D|H_A)} \frac{\Pr(H_0)_{prior}}{\Pr(H_A)_{prior}}. \tag{5.3}$$

This allows the comparison of two hypotheses, but requires that the prior odd ratio $\frac{\Pr(H_0)_{prior}}{\Pr(H_0)_{prior}}$ and the Bayesian factor $\frac{\Pr(D|H_0)}{\Pr(D|H_A)}$ can be estimated. One way to do this would be to calculate very many possible alternative hypotheses by varying the values of some interesting values of some parameter, and observing which values show low odds ratios for the null hypothesis. A lengthy example using coin tosses as data is presented in [76, p.249–251]. Another formulation of the Bayesian approach is [81]:

$$\Pr(H_0|D) = \frac{\Pr(D|H_0 \Pr(H_0)}{\Pr(D)}. \tag{5.4}$$

Here, too the probability of the null-hypothesis $\Pr(H_0)$ must be determined before any data is gathered.

A common argument in favour of Bayesian hypothesis evaluation is that it allows for evidence to strengthen either the null hypothesis or the alternative hypotheses, instead of merely evaluating the null hypothesis. The Bayesian approach offers also a method of cumulating the evidence across many studies. [76]

The use of Bayesian techniques for analysing experimental data has been at least as controversial as NHST [76]. Some researchers object to the Bayesian statistics on the grounds that it usually involve subjective appraisals because of the need to assign prior probabilities (which can be hard or impossible to estimate) to the hypotheses, and a probability to the data conditional on the alternative hypothesis [76, 98, 99]. Furthermore, the prior probabilities used in the Bayesian approach would change with time as knowledge about the studied phenomenon accumulates, which makes the conclusions from the Bayesian approach transient [99]. Such results would apply only for a short time.

It has been proposed to replace NHST with presentation and analysis of confidence intervals. Properly used, confidence intervals can be used for hypothesis testing [80, 83, 91, 98, 100, 106, 107, 108], but it offers more: Not just estimating, based on the data, whether the effect is or is not equal to zero, but the actual estimation of the effect itself [76, 82, 84, 91, 98, 99, 100, 105, 106]. A significant advantage of confidence intervals is that they provide a graphic signal of how much uncertainty there is in the data [107, 98]. The size of a confidence interval is determined by the same quantity that serves as the error term in ANOVA. Therefore confidence interval and the ANOVA lead to comparable conclusions [100]. The actual interpretation of the CIs is not as straightforward as it seems, but e.g. Cumming suggest a number of simple rules of thumb [109].

On the other hand, confidence intervals can be misinterpreted just as seriously as NHST [99, 110], and can be as arbitrary; what to conclude when the 95% confidence interval includes no treatment effects, but the 90% does [104]? It is often left unclear what the presented bars represent - standard errors, standard deviations, confidence intervals [88] or something else?

Confidence intervals tend to be rather large even when sample size is large enough for good statistical power [84, 105]. This is, in principle, good as the mere signal of significance communicates often a false sense of security or exactness. Cohen suspected that confidence intervals tend to be embarrassingly large and this is the primary reason for that they are not used [84]. Tryon noted that direct comparison of descriptive 95% confidence intervals constitutes a more rigorous standard of statistical difference than an ordinary t-test [77], although in must be remembered that “more rigorous” signifies also “less powerful”.

5.2 Sample size

It is clear from the above discussion on NHST that the sample size can be critical. Many studies are done with a too small sample size (see chapter 6). This is not surprising, as larger sample requires more work and acquiring subjects is not easy. The temptation to use a small sample is great, especially when it is easy to see from literature that many others are also doing that. However, as inadequate sample size can lead to wasted work, sample size calculations should be included when experiments are designed. Using a sample size large enough to achieve good power may be impossible in practise. In these cases it may still be preferable to carry the study out with a smaller sample size, since the alternative is a sample size of zero – that is, not doing the study at all [111]. However, even in these cases the sample size calculations will be useful as they will indicate the likely consequences of the selected sample.

The sample size is often determined by practical issues, such as time and money available. This is not guaranteed to lead to optimal or even adequate sample size. The required sample size depends on how the data will be analysed. The sample size is effected even by the decision of which variables and their formats are considered in the sample size decision [112]. For example, a dichotomous categorical left-right response format will require a larger sample than a seven-point response range measuring a continuous variable (e.g. naturalness of lighting) [112]. The information needed to determine accurately the good sample size is often not available preliminary information, making it more difficult to determine sample size (and later to justify it to reviewers) [111].

There are several criteria which can be used to estimate the required sample size. The simplest is just to refer to the practice in the field, or use what others have used. Another would be to find some recommendations and apply them. For example, the International Telecommunications Union recommend at least 15 observers for subjective assessment of video quality in multimedia applications [113]. The choice of the sample size for a scientific study should, however, be made according to more explicit criteria discussed below.

		Statistical significance test	
		Significant	Not significant
Practical importance of the result	Important	N good	N too small
	Not important	N too large	N good

Figure 5.3. Sample size and the practical importance of the results, according to Johnson [98].

5.2.1 External validity and representativeness

The sample needs to be independent random sample representing the target population. If the sample size or number of subjects is very small, external validity will become a concern [114]. If the subjects consist of five people, two men and three women, it is a good question to ask who they really represent, and to what extent are the results generalisable.

However, there have been studies with very small number of subjects, with results that are in general use for the entire human population. For example, the CIE colour system is based on the work of Wright and Guild in the late 1920's and early 1930's, who used ten and seven observers respectively [14, 115]. More extremely, the MacAdam ellipses characterising just perceptible colour differences were based on 25 000 observations –all by a single observer [14]. The results from this single observer are now assumed to represent the entire mankind (or at least the part of mankind with normal colour vision).

5.2.2 Statistical power

Statistical power is the probability to correctly reject the null-hypothesis in NHST, that is, to detect an effect when there is one. Much of the discussion of power in the psychological literature has focused on the question of how large sample is needed to have a high likelihood to obtain statistically significant results, i.e. to have adequate statistical power [105]. Some have challenged the use of statistical power analysis on the grounds that its primary purpose is to determine the sample size that is needed to obtain statistically significant evidence of an effect of a given magnitude [76, 105], and that this is pointless if statistical significance has little meaning [76] (see the above for the discussion on the statistical significance testing controversy). However, statistical results are generally more trustworthy when statistical power is high, both in avoiding Type I and Type II errors [105].

Some writers have also argued that power should not be so great that effects too small to be of interest will prove to be statistically significant [76, 98], because of added unnecessary work if nothing else. Johnson presented the four possible situations as seen in Figure 5.3. Sample size is good if it allows great enough statistical power to detect a meaningful difference, but no more. (It must be noted that Johnson presented this in an article criticizing the use of NHST).

A typical experiment involves a number of statistical comparisons. Even if the power of any single test is low, the power to detect some effect among multiple tests can easily be quite high [105, 116]. As statistical significance is often seen as the criterion for publication, the researchers can easily achieve some significant results and there is little incentive to choose sample size based on a serious consideration of statistical power or an effect size [105, 116]. This effects presupposes true effects but small power, i.e. the detected findings are a true positives. Performing multiple comparisons also increases the rate of false positives, and the significance criterion α is sometimes adjusted downwards to address this (see below).

Statistical power also depends on the statistical method used, and thus the decision on how the data will be analysed must be made before estimating the required sample size if statistical power is used as a criterion. Statistical power is a function of the effect size, significance criterion α, and the sample size N [92, 116], and thus the required sample size can be calculated when the power, significance criterion and the expected effect size are set. This can be done by calculating from the properties of the statistical methods being used, but there are published rules of thumb (e.g. [92]) and computer software (e.g. [117]) to do this.

Table 5.1. shows the sample sizes calculated by Cohen [92] and verified with G*Power 3.1.4 software [117] for three common statistical methods: The $t-$test to compare the difference between two independent means, The χ^2 (chi-squared) -test for goodness of fit (one way) or association in two-way contingency tables, and one-way analysis of variance (ANOVA). Cohen's results were generally similar to those calculated with the G*Power, but there were several cases where Cohen's sample size differed by one from that calculated with G*Power (perhaps due to different precision or rounding conventions). These are not reported in the table, but the two instances of greater difference between Cohen and G*Power are noted. The calculations are made assuming equal group sizes, and two-tailed tests. These sample sizes are per group or cell for the t-test and ANOVA, and the total sample size for the χ^2. If a medium effect was expected, and

Table 5.1. Sample sizes for α=0.05 and power of .80, according to Cohen [92] and calculated with G*Power software.

	Effect size		
Test	Small ($d = 0.2$)	Medium ($d = 0.5$)	Large ($d = 0.8$)
t-test (2 independent samples) per group	393	64	26
χ^2(df=1) total	785	87	26 (G*Power: 32)
χ^2(df=3) total	1090	121	44
χ^2(df=5) total	1293 (G*Power:1283)	143	51
ANOVA (2 groups) per group	393	63	26
ANOVA (5 groups) per group	240	39	16
ANOVA (7 groups) per group	195	32	13

the intention is to compare two groups with the t -test, 64 samples would be required from each of the groups for a total of 128. Similarly to use ANOVA to compare five groups to find an expected small effect would require a total sample of 1200. A 2x2 contingency table (with (2-1)(2-1)=1 degrees of freedom) analysed with the χ^2 would require the total sample size of 87 cases. These results show clearly that to investigate a subtle phenomenon where the effect size can be expected to be small, one needs a large sample to have a good chance of achieving statistical significance.

It could be argued that the power of 0.8 is too strict a criterion, although it is called criterion for what is “acceptable” [118, 119, 92]. After all, the lower the acceptable power, the lower will be the required sample size. If so, the question becomes how low can the power be set. Low power will not lead to false conclusions (as long as it is remembered that “no significance” is not a finding but an absence of one), but to increased chance of wasted work. The power of 0.5 can be regarded as an absolute minimum because an experiment with any less power would be more likely to not detect an effect than detect is assuming there is one. In other words, the experiment is less reliable in detecting an existing effect than tossing a fair coin. Table 5.3. shows the sample sizes associated with power of 0.5 (calculated with the G*Power software [117]). Very roughly, the sample size required for power 0.5 is about half of what is required for power of 0.8.

Not all recommend as large samples as this; Vanvooris et al. suggest 30

Table 5.2. Sample sizes for α=0.05 and power of .50, calculated with G*Power software[117].

Test	Effect size Small ($d = 0.2$)	 Medium ($d = 0.5$)	 Large ($d = 0.8$)
t-test (2 independent samples) per group	194	32	14
χ^2(df=1) total	385	43	16
χ^2(df=3) total	577	65	24
χ^2(df=5) total	700	78	28
ANOVA (2 groups) per group	193	32	14
ANOVA (5 groups) per group	130	22	10
ANOVA (7 groups) per group	109	19	8

per cell for power of 0.8, no less than 7 per cell for t-tests or ANOVA and at least 20 with no cell less than 5 for χ^2-tests. [120].

Post hoc power analysis

Power analysis is a powerful tool before the experiments to estimate the required sample size. It has also been used after the experiments for data analysis purposes, but this has been controversial. Calculations to estimate the achieved statistical power after the experiment have been suggested to be used as to interpreting tests with statistically non-significant results. This practice is been widely criticised as inappropriate [105, 108], although it can be useful if correctly done [121].

There are two common approaches when a non-rejected null hypothesis occurs. The first is to compute probability of rejecting the null hypothesis, assuming that the true effect and variability are accurately estimated by their observed values [108]. In this approach, the lack of statistical significance despite a high observed statistical power (calculated *post hoc* from the data) is interpreted as evidence for the null hypothesis being true [108]. This reasoning fails, because the p -values determine the observed power. Therefore a non-significant p-value automatically correspond to low observed power [98, 108, 121, 122]. The estimate of the achieved power is also often far from precise, and tends to over-estimate the power [121].

The second approach is to calculate what the true difference would have needed to be to achieve a given statistical power, in order to estimate the effect size that would have probably been detected [108]. This is then used for an upper bound on the true effect size. A variation is to calculate the power at a specified effect size, chosen so that the effect of this size signifies a practically important result [108, 121]. The logic behind both methods is that the higher the estimated power is for detecting meaningful departures from the null, the stronger the evidence is taken to be for nature to be near the null when the null is not rejected [108]. The disadvantage is that it can be difficult to estimate the practical importance of a standardized effect size measure [121].

If it is desired to verify that the effect is close to null, a more fruitful approach than power analysis is to use various methods for equality testing [77], or to estimate the effect sizes and their confidence intervals [98, 105, 121].

5.2.3 Parameter estimation and confidence intervals

When estimation of the magnitude of a parameter is important, it is not only the estimate itself what is important, but also the width of a confidence interval for the parameter in question. When the parameter is estimated precisely, the standard error will be small, and thus the width of the confidence interval will also be small. A major advantage of sample size planning for accuracy is that sample size formulas for narrow confidence intervals can be much less dependent on the actual value of the population parameter in question than are sample size formulas for power. What is considered acceptable precision depends on the situation. If the predicted confidence interval is too wide, the simplest solution (other than decreasing the level of confidence below 95%) is to obtain a larger sample. [105]

Goodman and Berlin derived rule of thumb equations for predicting the range of 95 % confidence interval [122]:

$$\begin{aligned} CI_{95,pr} &= \Delta_{obs} \pm 0.7\delta_{80} && (5.5) \\ &= \Delta_{obs} \pm 0.6\delta_{90} && (5.6) \end{aligned}$$

where

$CI_{95,pr}$ is the predicted 95 % confidence interval

Δ_{obs} is the observed difference

δ_{80} is the true difference for which there is 0.8 power

δ_{90} is the true difference for which there is 0.9 power.

However, Goodman and Berlin did not suggest that power calculations should be replaced with these equations, because if this was done, the tendency would be to make the precision exactly equal to the difference of interest, resulting to a study with power of 0.5 instead of the 0.8 usually recommended [122]. They also warned that the equations work only approximately when the sample size is small (less than 20). It should also remembered that these equations are for *predicting* the confidence interval, not calculating them after data is obtained.

From this it is easy to see why a sample size determined by the criteria of statistical power may not be enough for precise estimation. For example, The sample size is chosen so that the experiment had power of 0.8 with a medium-size effect (Cohen's $d = 0.5$), giving a sample of 64 per group (Table 5.1). The predicted confidence interval expressed in the same scale as the effect size would be $D_{obs} \pm 0.7 \cdot 0.5 = D_{obs} \pm 0.35$. In other words, if the detected effect size was approximately the expected 0.5, the 95 % confidence interval would run from effect size of 0.15 to 0.85. It is clear that there was an effect and to which direction it was, but the confidence interval is still consistent for a small (Cohen's $d = 0.2$), medium $(d = 0.5)$ or large $(d = 0.8)$ effects. Let us assume that the true effect was medium and it was desirable to set the 95 % confidence interval so that a large or a small effect can be discounted. The $CI_{95,pr}$ is thus set to range from 0.25 to 0.75. Therefore, a $\Delta_{80}$of $0.25/0.7 = 0.3571$ would be necessary. A calculation with the G*Power software suggest the sample size of 98 per group, which is more than 50 % increase from the figure of 64 per group calculated by the statistical power criterion.

One method of determining sample size based on parameter estimation is to specify the acceptable margins of error for the most vital parameters to be estimated [112]. The required sample size to obtain an accurate estimate is often larger than the sample size necessary for adequate power, but the reverse can also be true, depending primarily on the size of effect to be detected [105]. The sample size required for accurate parameter estimation depends on the method being used. Bartlett et. al presented [112] equations designed originally by Cochran. For estimating the mean of continuous data they suggested

$$N = \frac{t^2 \cdot s^2}{e_{acc}^2}, \tag{5.7}$$

where

t is the value of the t-statistics for the selected criterion for significance (1.960 for $\alpha = 0.05$ and 2.576 for $\alpha = 0.01$ for large samples and two-tailed tests)

s is the estimate of the standard deviation in the population

e_{acc} is the acceptable margin of error for the mean

The application of equation 5.7. has some practical limitations. First, the value of t depends on the sample size unless the sample size is more than about 120, which may make the process iterative. The second difficulty is that an estimate of the standard deviation in the population in question must be available. In some cases, a pilot study or the published research literature may provide a workable estimate of the standard deviation. Otherwise the researcher is forced to rely on a estimate or guess based on the logic or mathematical structure of the scale. For discrete response range, Bartlett et al. suggested that the standard deviation can be estimated with the ratio of number of response options to the number of standard deviations that would include all possible values in the range (they suggested six). [112]

Based on these calculations, Bartlett et al presented a table for determining the sample size. They were writing for organizational research in mind, so they accounted for the possibility that the total target population can be fairly small. The largest population in the table of Bartlett et al. was 10000, for which they suggested the sample size of 119 for continuous data and 370 for categorical data, assuming αlevel of 0.05 and 3 % and 5 % acceptable margins of error for continuous and categorical data respectively [112].

For example, the researcher has decided that $\alpha = 0.05$ will be an adequate criterion for significance, and that the acceptable margin of error for the estimate of the mean is 0.25 (corresponding to e.g. 5% error on a five-point response range), and he estimates the standard deviation to be $s = \frac{5}{6} \approx 0.83$ for the five point response range he is using. If he used equation 5.7 to calculate the sample size, he would conclude that the needed sample size is

$$N = \frac{1.96^2 \cdot 0.83^2}{0.25^2} \approx 42 \qquad (5.8)$$

This sample size is clearly smaller than what power analysis or the method by Goodman and Berlin suggest. The result is also quite sensitive for the acceptable margin of error. If 10% error was allowed in the previous calculation, the resulting sample size would be

$$N = \frac{1.96^2 \cdot 0.83^2}{0.5^2} \approx 11, \tag{5.9}$$

which is unlikely to achieve reasonable statistical power.

5.3 Multiple comparisons

When the study consist of making several comparisons, each having the critical parameter of significance α, each comparison presents an opportunity for committing the Type I error (a false positive), increasing the overall probability that at least one Type I error will be committed. Assuming that the null hypothesis is true, the probability of making a right choice when rejecting the null hypothesis is $(1 - \alpha)$. When k comparisons are made (assuming that the null hypothesis is always true, and that the comparisons are independent) , the probability that all the decisions were right (i.e. no Type I errors were committed) is $(1 - \alpha)^k$, and thus the probability of at least one Type I error is $(1 - (1 - \alpha)^k$. For example, if ten comparisons were made at αlevel of 0.05, the probability that at least one of them involved a Type I error is $1 - 0.95^{10} \approx 0.4$.

To address this, several statistical procedures have been developed. One of the most common is the Bonferroni correction, which is used to adjust the α-criteria for statistical significance downwards, the usual suggestion being to use α/k for the significance criterion instead of α. There are, however, other *post-hoc* methods such as Tukey's or Duncan's test, for which the reader is encouraged to consult a good statistics textbook.

Whether to adjust the α values for multiple comparisons is controversial. While doing so achieves its aim of reducing the rate of Type I errors, it at the same time increases the rate of Type II errors (false negatives). Many consider the need for the adjustment as a fixed rule. For example, Kuzon et al. included using unmodified t-tests for multiple comparisons in their list of "seven deadly statistical sins" [119]. Others, however, criticises its application [123, 124, 125], or at least mechanistic application [111], because the using the adjustment leads to loss of power and therefore publication bias and incorrect reporting of no effect when statistical significance is not obtained [123, 125]. The Bonferroni correction has also been criticised from mathematical principles; it guards against making even one Type I error. However, when there are many comparisons, even if it is likely that a small number of Type I errors emerges, it is very unlikely that many of them occur. It is thus often more useful to look for coherent patterns of results, and when one emerges, the results reinforce each other rather than negate each other [123, 124].

It has also been pointed out that with correction for multiple comparisons, the result of all the individual comparisons depend not only on the data and on the critical level of significance, but also on the number of questions asked [103]. The choice on how to apply the adjustment is also arbitrary; should the correction be calculated for all comparisons made in a single table, for all comparisons in the entire paper, for all comparisons in the entire issue or volume of a journal (or all issues so far published), or all comparisons the researcher has done during his career [103, 124]? While the latter alternatives were clearly presented as an *reductio ad absurdum,* the first choice between the first options are not so clear. There are always some leeway for the researcher to manipulate how he presents his data, in how many tables he divides his findings. The Bonferonni correction can also be made less severe by presenting less comparisons than were actually made by omitting some or all of the variables.

When a multiple comparison procedure will be used in data analysis, sample size planning should take this into account. Without such consideration, sample size will likely be too small [105]. To restore the statistical power lost due to the adjustment requires much larger sample size than if the multiple comparisons were not compensated for. For example, suppose an experiment involved two lighting conditions and each was evaluated with a questionnaire with eight items. The results are compared with a two-tail independent samples t-test. If it was decided to keep the level of α experiment-wise 0.05 and use the Bonferonni correction, the comparisons would have use the criteria for significance of $\frac{0.05}{8} = 0.00625$. The required sample size (calculated with the G*Power software [117]) to achieve power of 0.8 for a medium effect is 64 per group if no Bonferonni correction were made, but 105 if the correction was used.

Whether or not to use the correction for multiple comparisons depends on the situation. The crucial question when considering applying the correction for multiple comparisons is what is more critical, Type I or Type II errors. If it is critical to avoid committing Type I errors, even at the cost of low power and high possibility of Type II errors, then the correction needs to be used.

The possible effect of multiple comparisons needs to be considered when interpreting the results, whether of one's own work or those of others. When there are many comparisons and few statistically significant results, the real significance of those results may be well doubted. But when there are many comparisons and many significant results, the presence of one or few false positives are usually not fatal to the general conclusions. The simplest answer to multiple comparisons is to avoid using too many variables and therefore comparisons [126]. Garamszegi suggests that instead of using Bonferonni correction, effect sizes and confidence intervals should be presented [123].

5.4 Parametric or non-parametric methods?

Most commonly used statistical methods make the assumption that the data is independently drawn from the same population which approximately follows a statistical distribution which can be described with prespecified parameters. These methods are called parametric methods. For many of them, there exist comparable procedures which do not make the assumption on normal population, called non-parametric methods. The non-parametric methods are usually based on transforming the scores into rankings, and comparing these rankings.

In most cases, estimating parameters is interesting, making parametric methods are natural choice. Non-parametric methods can be used for estimating parameters, but they are better suited for null hypothesis significance testing [127]. Non-parametric methods are often applied only in situations in which the assumptions underlying the parametric methods are clearly violated, and doubts arise whether the robustness of the usual parametric method is sufficient to allow its use [128]. Parametric methods are, however, known to be quite robust as long as the violations on the underlying assumptions are not extreme .

Concern about he normality of the distribution is especially common reason for the choice of a non-parametric method, and it is not rare to see sentences like "Our data were not normally distributed, so we used non-parametric methods" in the literature [114, 127]. However, for the parametric tests ($t-$tests, ANOVAs and so on), the assumption of normality applies to the distribution of means, not to the distribution of the data [114, 127], and The Central Limit Theorem assures that the means are approximately normally distributed when the sample size is not very small [114, 127] (greater than 5 or 10 per group [114]).

Some argue for wider use of non-parametric methods because they are said to be free from assumptions about the distributions, but this is not entirely true. For example, the MW-test, which is a non-parametric analogue to the t-test, test actually the hypothesis that the two distributions are identical, not that they have the same mean [127]. When the MW-test is used to compare means (or medians), it is usually tacitly assumed that the result of "statistically significant" is entirely due to difference in the central tendency (means or medians); the test, however does not reveal whether the result was due to differences in central tendency, shape, variance and so on. It is possible that two distributions have equal means (or medians), but the MW-test returns a statistical significant difference. This is not error, since what was tested was "are these distributions identical", not "are the means of these distributions equal". It is often not clear why the assumption that the two distiributions have similar shape and differ only by location is more justifiable than the approximate normality assumed by parametric tests [127].

Robustness against false conclusions is different from statistical power, however. Parametric methods are said to have better statistical power than the corresponding non-parametric tests [75, 114, 128]. However, there are many studies showing that the power of non-parametric tests is not much inferior, and can be clearly superior when the distribution of the data is not normal. Blair and Higgins used Monte Carlo methods to compare the relative power of the paired samples t-test and Wilcoxon's signed-ranks test under different population shapes [128]. Which of the tests was more powerful depended on the situation, but the power advantages of the t-test under normal theory were small while the $t-$test was often much less powerful in non-normal situations. They concluded that the claim that parametric tests are more powerful than non-parametric tests is not justified for these two tests [128]. Nanna et al. compared t-test and Wilcoxon ranking sum -tests with a Functional Independence measure, and found that the Wilcoxon rank-sum test outperformed the t-test for almost every sample size and alpha level examined [129]. Gregoire and Driver found no clear-cut superiority in performance between parametric and non-parametric methods when using simulated Likert-type data [130]. Winter et al. compared performance of the t-test and MW-test with an extensive variety of simulated five-point Likert-type data, and found that the power differences between tests were minor and concluded that researchers do not have to worry about the rate of occurrence of false positives when making the choice between these methods [27]. Schmider et al. used Monte Carlo methods to investigate how robust ANOVA is concerning violation of the normality assumption, and found that the empirical rates for Type I and Type II errors remain constant under different non-normal distributions, and concluded that ANOVA can be regarded robust [131] if the commonly given advice to use sample size of at least 25 participants per condition is followed [131].

Some claim that non-parametric methods should be used when the sample size is small. However, there is no evidence that non-parametric tests are more appropriate than parametric tests when sample sizes get smaller [114]. Blair and Higgins found that when Wilcoxon's test was more powerful, the magnitude of the power advantage often increased with increases in sample size [128].

Thus, the literature presents no overwhelming reason to categorically prefer parametric or non-parametric methods. The choice should be based on what is being asked from the data. If a parameter is being estimated, then parametric methods are a natural choice. There is little risk of coming to the wrong conclusion [114, 127], and the most commonly cited reason to abstain from using them, concern for non-normality, is usually not a factor if the sample size is sufficient even for a large effect. In other situations, such as when the interesting question concerns rankings, non-parametric methods may be more powerful.

5.5 Ordinal or interval data? Does it matter?

Stevens introduced in 1946 the concept of a relationship between a measurement format and the statistics method that should be used [132]. He classified psychological response formats into four classes is presented in Table 5.3. This classification was a significant contribution to psychophysical and measurement theories, and the classification became widely accepted, although criticized by statisticians [133, 134, 135]. The theory has been developed further since Stevens' original work, and some of his concepts gain their true value only with the later developments [136].

The simplest level of measurement is nominal; the response options only determine unordered categories. If one option is labelled 1 and another 4, it would be perfectly logical to switch them around in a nominal scale. The options could be as well designated with alphabets. An example question of this nature would be the sex of the respondent; the choice to label is completely arbitrary – one could as well cite the “ladies first” -principle and assign the option “Female” number 1, as the number of X-chromosomes and assign the “Male” option number 1. The only rule is not to assign the same number for different options, or different number for the same options. The results of such data would most naturally be reported as the number of each case (e.g. 47 participants, 22 male and 25 female, participated in the experiment), or percentage or other proportion of the total number of cases. The central tendency is expressed as the mode, i.e. the most often chosen option.

An ordinal type measurement is one where the options are categories, and these categories can be arranged in a logically progressing, monotonically increasing order. A typical Likert-type response format where the options range from “Strongly disagree” to “Neither agree nor disagree” to “Strongly agree” is an example of an ordinal type (however, see below). However, the interval between the options is not defined. The central tendency of ordinal data can be expressed with the median.

The interval type includes an operation for determining equality of intervals, for determining greater or less, and for determining equality (not greater and not less) [132]. Because the intervals are now defined, the central tendency can be expressed with the mean.

Ratio-type also includes the operation for determining the equality of ratios, and requires a true zero point (i.e. one that cannot be shifted by adding a constant) [132].

It must be noted that the levels of measurement are increasingly stricter in their criteria, and the stronger level includes all the properties of the lower. Thus, the column listing in the Table 5.3. the basic operations needed to create each type of scale is cumulative: all operations in the preceding rows must also apply. Similarly, all the statistical methods listed in the last column are admissible for ratio data, and interval data can be evaluated with median and percentiles or mode and so on [132].

For each data type, Stevens presented a group of admissible mathematical transformations which share the property of describing the measured property equally well, i.e. the findings or conclusions do not depend on the transformation used [132, 137]. Stevens also specified the appropriate or “permissible” statistical measures for use with each measurement format. His criterion for appropriateness was that an appropriate statistical method remains invariant under the admissible transformations [132] (although it can be problematic to define the precise meaning of in-

Table 5.3. The four levels of measurement, according to Stevens [132].

Type	Basic Empirical Operations	Mathematical Group Structure	Permissive Statistics (invariate)
Nominal	Determination of equality	*Permutation group* $x' = f(x)$; $f(x)$ means any one-to-one substitution.	Number of cases Mode Contingency correlation
Ordinal	Determination of greater or less	*Isotonic group* $x' = f(x)$; $f(x)$ means any monotonic increasing function.	Median Percentiles
Interval	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$; a and b are constants.	Mean Standard deviation Rank-order correlation Product-Moment correlation
Ratio	Determination of equality of rations	*Similarity group* *x'=ax;* a is a constant.	Coefficient of variation

Table 5.4. Examples of some admissible transformations allowed for ordinal data.

	x_i	y_i	x_i^2	y_i^2	$\sqrt{x_i}$	$\sqrt{y_i}$	$\ln x_i$	$\ln y_i$	2^{x_i}	2^{y_i}
	1	1	1	1	1	1	0	0	2	2
	2	1	4	1	1.4	1	0.7	0	4	2
	3	3	9	9	1.7	1.7	1.1	1.1	8	8
	4	5	16	25	2	2.2	1.4	1.6	16	32
	5	5	25	25	2.2	2.2	1.6	1.6	32	32
Mean	3	3	11	12.2	1.5	1.4	1.0	0.9	12.4	15.4
Median	3	3	9	9	1.7	1.7	1.1	1.1	8	8

variant [138]). Table 5.4. shows an example of transformations, under which the statics permissible for ordinal data remain invariant, but those for interval data do not. A hypothetical question with five response options labelled with integers from one to five is presented to two groups (x and y) of five people each. The respondents in group x answered to the five questions with 1, 2 ,3, 4, 5 while the respondents in group y answered with 1, 1, 3, 5, 5. It is easily seen that both the means and the medians of these responses are equal. A number of transformations are then presented (all are monotonically increasing when the possible arguments are positive as in this case, this allowable for ordinal data). As seen from the Table 5.4., not only is the equality of the means not preserved, but which of the two means is greater depends on the transformation performed. Medians, on the other hand, remain equal.

Thus, only non-parametric methods are invariant under the admissible transformations, and therefore permissible with nominal and ordinal data. Parametric methods are permissible for interval and ratio formats only. This rule is very often presented as a strict "do so" way in textbooks and literature written to guide the researcher (e.g. [29, 119, 1, 139, 140]). This is often done in a misleading way, because the critical qualitative conditions that distinguish different scale structures are often omitted in the discussion [22, 129, 135, 141]. Especially, the two middle columns of the table 5.3 and the whole concept of allowable transformations are omitted. This is a very serious omission, since the invariance under the allowable transformations was Steven's *sole criterion* for determining the permissible statistical methods. The term "permissible" is then misunderstood to the effect that a method that is not permissible must be incorrect, wrong or illegitimate for some arcane statistical reason, rather than just not invariant under certain transformations [136, 137, 141]. The result

a) The lighting in this room is *stimulating*

Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
□	□	□	□	□	□	□

b) The lighting in this room is *stimulating* (a=strongly disagree, g=strongly agree)

a b c d e f g

c) The lighting in this room is *stimulating*

strongly disagree 1 2 3 4 5 6 7 strongly agree

d) The lighting in this room is *stimulating*

strongly disagree -3 -2 -1 0 1 2 3 strongly agree

e) The lighting in this room is *stimulating*

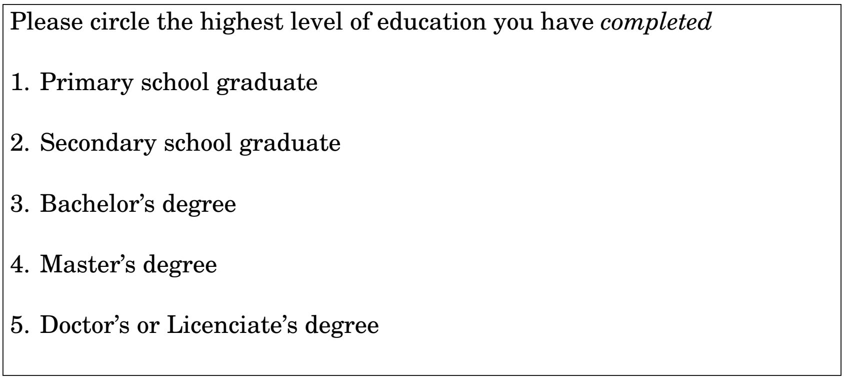
strongly disagree ———————— strongly agree

Figure 5.4. Increasingly interval-like Likert-type response formats.

is that the issue is often presented as more clear-cut than it actually is [133, 141].

Determining the level of measurement for a format is not always easy [133, 142, 143]; even the supposedly arbitrary symbols for nominal categories may in fact contain relevant data [133]. Depending on the presentation and context, a response format may be more or less interval- or ordinal-like even if the basic question and task remain the same. There are no agreed-upon rules for determining when a particular response format is ordinal, less than ordinal, or more than ordinal [143]. An example of this is presented in the Figure 5.4. The first format offers a choice of ticking a box, and the options are clearly categories, making the item clearly ordinal. The format in Figure 5.4. b) is somewhat more vague having some characteristics of an interval-type measure. While the options a–g still represent categories, the stimulus information (a=strongly disagree, g=strongly agree) and lack of labelling make it somewhat more interval-like than the first response format. The response format in the Figure 5.4. c) replaces the alphabets with integers. This situation is somewhat unclear. Some could argue that it is no different from version b), as the only change is the symbols used to label the alternatives. Others might see the version as essentially interval-like, where the numerical labels serve as milestones on a continuous range. The version d) is similar to c) but has shifted the range so that the labels reinforce the bipolar nature of the task. The visual analogue format in Figure 5.4. e) is the other extreme; such response format is clearly of the interval-type, and would require arbitrary rules for grouping to form categories (although, some argue that even visual analogue scales produce ordinal data, [140, 144], claiming that the assessment is highly subjective and the result cannot thus be interpreted as numerical [144]). The issue is made even more unclear in that the version the subjects see is not necessarily the one the researcher assumes in his analysis. For example, the subjects are presented with the version a), but the researcher codes the answers as the questionnaire was version c) or d). The data type is not thus an absolute category, which can be assigned by e.g. being excluded from the type above it, but it depends on the context [133, 138, 142]. Jacoby argued that it is not the nature of the variable, or the data, but the researcher, which decides which level of measurement is appropriate [142]. It has even been argued that we do not in fact know the set of permissible transformations for most of the scales in behavioural sciences [141].

Sometimes it is clear that ordinal categories, even if labelled with numbers, do not lend themselves well to representation as means. Consider, for example, the background question on the subjects' education presented in the Figure 5.5 . The range is ordinal, as a clear order can be established. However, few if any researchers would report the subjects' attained education with e.g. "The subjects' ($N = 64$) mean education level was 2.2561 ($s = 0.5525$)...". The interpretation of the calculated mean is also problematic. If subjects included a post-graduate student who has completed his Master's degree, and a first-year student who has not yet completed his Bachelor's degree, very few would say that it is equal to having one who has only finished the primary school and one with a doctoral degree, or to two persons with a bachelor's degree. (Of course, the standard deviations would be different, but that is a different matter than interpreting the mean). A similar argument has been made for Likert-type response formats; "A strongly agree and an agree does not make two agree-and-halfs [29]". This can be especially significant in a "love it or hate it" -situation, where truly indifferent opinions are rare. The problem of interpretation, however, can be equally true even if the central tendency is expressed



Please circle the highest level of education you have *completed*

1. Primary school graduate
2. Secondary school graduate
3. Bachelor's degree
4. Master's degree
5. Doctor's or Licenciate's degree

Figure 5.5. Numbers, where calculating mean would be nonsensical.

with median; with even number of cases the median can be "something and half" even if there is no such option in the response format. Therefore, the problem of interpreting the number describing the central tendency is not solely a level of measurement problem, nor it is entirely avoided by observing the principle of permissible statistics.

Even if the score 2.2561 does not have a real-world reference, it can be argued that it still gives some information about the average attained education level of the subject population, and the question whether it is different from another sample population whose mean score was 2.6131 can be interesting and relevant, and a satisfactory answer can be found via statistical methods.

5.5.1 Different philosophical approaches

The controversy about measurements and statistic has continued for over seven decades, and shows no sign of reaching a consensus. Even the basic arguments remain the same. Michell argued that the disagreement lies deeper than the arguments presented [145]. According to him there are at least three philosophies of science that yield distinct positions on measurement/statistical issues: representational theory, operational theory, and classical theory.

The representational theory is based on the idea that numbers are used in measurement to represent empirical relations between the objects being measured [145]. The numbers and the number system must represent empirical facts that exist independently of their representation [146]. Any conclusions reached via mathematical operation of these numbers must be wholly implied by the empirical data itself, and cannot depend on the numbers assigned [145]. Therefore, a researcher subscribing to this philosophy of measurements would agree with Steven's criterion of invariance under the allowable transformations for statistical methods he would consider permissible.

The operational theory sees the measurement simply as an operation that produces numbers or numerals, and the meaning of these numbers are conferred by the operation [145]. This differs from the representational view in that the operational theory does not require the numbers to point to any empirical structure [145]. Thus, the relationships between the numbers are viewed as a scientific endpoint themselves [146]. Thus, any consistent operation which reveals scientifically interesting relationships between the numbers is legitimate to an operationalist. He would not see any reason to limit himself to non-parametric methods for ordinal data.

To the classical theory, a measurement is an assessment of quantity, i.e. how much of a given attribute some object possesses. To be measurable, a quantity must be quantitative. The number represents quantity, and their relationships are not thus assigned but discovered in a measurement. The classical theory only requires observational evidence supporting the hypothesis that the attribute to be measured is quantitative, but no limitation is put on the form this evidence must take. The classical theory sees the measurement results always as real numbers, and thus any valid numerical argument form may be applied to them. Therefore, this theory, too, rejects Steven's classification and restrictions to what is permissible and what is not. [145]

In similar vein, Gaito observed that the controversy seems to be due to a point of view difference between measurement theory and statistical theory [134]. According to him, the measurement theory requires that the numbers used must be meaningful, relative to the characteristics of concern. For statistical point of view, differences and relatedness of numbers is what is important. The meaning of numbers does not enter the picture because, statistics do not make assumptions on how the numbers were obtained, only on their distributions, Gaito argued. [134]

5.5.2 Arguments against using parametric statistical methods for ordinal data and their counterarguments

Many consider using parametric methods for ordinal data wrong, illegitimate [29, 136, 140, 147], or unacceptable [144, 148]. Kuzon et al. went as far as consider it the first of their "seven deadly sins of statistical analysis"

[119], and Svensson argued that there is ethical considerations against it [140]. There are five commonly used arguments against using parametric methods with ordinal data:

1. The admissible transformations define the permissible statistical methods, and transformations admissible for ordinal data rule out parametric methods.

2. The intervals between the categories are unknown and cannot be presumed to be equal as assumed in parametric methods.

3. Since only the order of categories is known, the intervals between them are undefined, which makes basic mathematical operations such as adding between them invalid.

4. Ordinal data cannot be normally distributed, which rules out using any statistical method that assumes normality.

5. The integers are only labels, and sums and means derived from them are meaningless and do not relate to anything real.

Each of these arguments and their counterarguments will be discussed below

The admissible transformations define the permissible statistical methods, and transformations admissible for ordinal data rule out parametric methods.

The concept of admissible transformations rules out the use of parametric methods if strictly adhered to [132, 136, 137]. The ordinal group admits any order-preserving (i.e. monotonically increasing) transformation, which allows transformations which do not preserve the equality or order of means. The initial creation of the system for encoding the responses involves arbitrary decision by the experimenter, and may contain implicit transformations. As Fotios noted, the process where the response format forces the original responses into a limited set of categories is tantamount to subjecting the original response to a non-linear transformation [8]. Thus, utilising only admissible statistical methods ensures that conclusions about the measured property do not depend on implicit transformations or arbitrarily chosen numerical values [136, 137, 146]. The argument is easy to demonstrate with calculations, as the ones presented in the table 5.4. Townsend and Ashby demonstrated also a number of examples where statistical significance depend on the transformations [136], using this to argue that only statistical methods immune for this effect (i.e. those Stevens called permissible) can be used.

The most common counterargument to this is to disagree with Stevens concept of permissible statistics [133, 134]. The requirement that only transformation-invariant statistical method is permissible may be too strict for practical data [133]. It is undeniably true that certain transformations allowable for ordinal data will distort the data so that parametric methods can become misleading, or their result depend on the arbitrary transformations made. However, it can be argued that this is a problem associated with the transformations, not the parametric statistical methods. Since some transformations can make the data more amenable for good analysis it would be a loss if one refrained from using them unless for a very weighty reason [133].

The concept of allowable transformations limiting the permissible statistics is not always used consistently. It is used as an argument against using parametric methods for ordinal data, but ignored it otherwise. Strictly applied, it would preclude using parametric methods if, for example, a logarithm transformation of the results is used for analysis. Is the fact that the logarithm of the means is generally not equal to the mean of the logarithms really a sufficient reason to prohibit comparing either means with parametric methods? What if there are theoretical reasons to suspect logarithm-like relationships?

Furthermore, Michell argued that to report the median or mean of a set of measures is to report a fact about them. Does it serve science to ban such reporting when it may contain otherwise interesting information [145]?

The intervals between the categories are unknown and cannot be presumed to be equal as assumed in parametric methods.

Some (e.g. [29, 119, 129, 136, 148]) object the use of non-parametric methods because the intervals between the categories is not known and cannot be presumed to be uniform. Therefore, parametric methods are invalid for such data.

There are two principal types of counterarguments for thus argument. One is to show that the assumption that the intervals are approximately uniform is reasonable [133], or empirically vindicated [22, 28]. Studies on the optimal number of response categories involve linear rescaling procedures [9, 26, 35]. The results of these studies also support the view that the assumption of approximately uniform intervals is not unreasonable, at least when there are more than five response category options. Careful choice of the response format and labelling can also be used to signal that the categories are approximately equidistant [23].

The other typical counterargument is that the parametric statistical methods do not require that the intervals are uniform, only that the distribution is approximately normal [114, 134].

The intervals between the categories are undefined, thus basic mathematical operations depending on these intervals are invalid

A stronger version of the previous argument is that the intervals between the categories are not just unknown, but undefined; since the categories could be as well be labelled with "a, b, c..." as with "1, 2, 3...", using integers as labels does not mean that there is anything in the structure in between the categories. In much of the literature, no distinction is made between this and the previous argument. However, it is very different mathematically whether a quantity is undefined, or defined but unknown. If the intervals are merely unknown, one could, at least in principle, overcome this limitation with trying to estimate or set upper and lower bounds for the unknown values.

The lack of defined intervals means that the basic mathematical operations (e.g. addition, multiplication or division) for them are not valid [29, 119, 140, 149]. Therefore, sum-scores, the mean value, standard deviation and calculation of differences for description of change in score do not have an interpretable meaning with ordinal data [29, 138, 140, 144, 148], and thus parametric methods test for nonsensical hypotheses and thus cannot be used.

This argument can be countered with the observation made in the previous point, that there are empirical reasons to think that the intervals are, in practice, defined and approximately uniform.

Ordinal data cannot be normally distributed, which rules out using any statistical method that assumes normality.

Some argue that since ordinal data is discrete with unknown intervals and restricted range, it cannot be considered normally distributed [129, 150], and thus any statistical method assuming normality should not be used. Others point out that there is no reason to assume the existence of an underlying interval scale when a frequency distribution is found that approximates the normal distribution [146]. This argument is presented even by some who otherwise dismiss level of measurement from consideration in choosing between parametric and non-parametric tests [129].

The counterargument to this is that the data can usually be considered as a sufficiently close approximation of the normal, and that the parametric methods are robust enough to reduce this to non-issue [22, 28, 114, 134]. Gaito claimed that if the data follows a normal distribution, then the data would be of interval scale nature because the intervals between any data points are known (in terms of probabilities, i.e., areas under the curve) [134] However, Thomas proved false at least in the strict sense the premise “if normality, then interval” [135]. Scores obtained by summing the results from individual items tend to be more normal-like distribution than the individual items [27].

The integers are labels only, and thus their sums or means are meaningless and do not relate to anything real.

The categories are labelled with any monotonically progressive way. The integers used to label the groups are merely labels, thus while a category can be labelled “2”, this does not signify that there is twice as much some quantity as in the category labelled “1”. The labelling system is arbitrary from the mathematical point of view under the allowable transformations (not, of course, from what it signals to the respondent), so it would be equally valid to label the groups “1, 2, 3, 4, 5” as it was “1, 10, 501, 501.4, 10387”, provided that only non-parametric methods were used to analyse this data.

Consider, for example, an item where glare is evaluated with five options (0=No glare, 1=slight glare, 2=moderate glare, 3=severe glare, 4=very severe glare). These are clearly categories, and they can be ordered so this is an ordinal scale. It is clear that slight glare is less than moderate or severe glare. In this coding, the arithmetic difference between any two options next to each other is always one, but it is impossible to say that the difference between no glare and slight glare is equal to that between moderate and severe glare. Mathematically 2+3=5, but what is the meaning of “moderate glare added to severe glare?” According to this argument, only the labelled points are meaningful. We can calculate that the arithmetic mean of 2 and 3 is 2.5, but “the sum of moderate glare and severe glare, divided by two” has again no meaning. The essential question is, what does such mean actually measure [149]?

The counterargument is that the meaning of the numbers or the inferences made from them is a different issue. Statistical methods can still be used to verify if it is justified to say, for example, that the result is different from being exactly 2, even if there is nothing that the score 2.5 represents. Refraining from using parametric statistics does not remove this problem as pointed above, since the median of 2 and 3 is also 2.5.

The terms "meaningful" and "meaningless" are also less than clear in many cases [133, 143]. Even if the score 2.5 has in itself no significance, in the context of the results it can still give information about how the responses were distributed. Even seemingly meaningless statements may contain scientifically interesting information [133].

It can also be argued that the labels help to define the response range, and equally placed integers guide the respondent towards thinking the response options more or less equidistant, which would also address that argument against using parametric methods. The difference between labels can thus matter and the choice of labels not completely arbitrary.

5.5.3 Arguments for allowing using parametric statistical methods for ordinal data and their counterarguments

There are many, however, who do not object to using parametric methods with ordinal data. Following arguments can be found from literature for use of parametric statistical methods:

1. Parametric methods are sufficiently robust to handle ordinal data.

2. The questionnaire data actually is, or can be considered as interval data.

3. The level of measurement is not an assumption made on the parametric methods.

4. Statistical methods are indifferent to how the numbers were obtained, the meaning of the numbers is outside statistics.

5. Parametric methods have been widely used to analyse ordinal data for decades with fruitful results. Thus the practice of doing so is supported empirically.

Each of them will be discussed more closely below along with counterarguments presented against them.

Parametric methods are sufficiently robust to handle ordinal data.

The proponents of this view admit that the arguments against the use of parametric statistics with ordinal data may be strictly speaking true, but they are irrelevant in practice because of the robustness of parametric tests [22, 27, 114, 131, 134]. In other words, the probability statements produced by the use of statistical methods remain valid even if the underlying assumptions are violated [129]. There is considerable evidence, both from simulation and real data, that this position is justified [27, 114, 129, 131, 134, 141, 151], although the assumption has been challenged with contrary examples [151]. This robustness has also been demonstrated for correlation methods [114, 151].

Robustness is, however, different from statistical power. Several studies have found that with non-normal distributions, non-parametric methods can outperform their parametric counterparts [129, 143], although others state that the parametric methods would be better as they tend to be more powerful [28, 75, 114, 152]. The difference between these may be due to the distributions analysed; some investigated relatively normal distributions, whereas others used distributions of real data which were considerably skewed [27].

This argument has been countered by demonstrating either simulated or arithmetic cases where the robustness fails. Townsend and Ashby, for example, used a case where the ratio of two temperatures is equal if the temperatures were expressed in Fahrenheit, but resulted into a division by zero if the same temperatures were converted into Celcius [136]. They illustrated also that whether the test of significance between two groups of respondents remain invariant can only found out empirically, not with a priori simulation or transformations.

Kahler studied whether the choice of method affects the conclusion in a quality of life measure [147]. She found that applying simple parametric methods to ordered categorical data can led to different results from those obtained with non-parametric methods, although by this she meant that the cases where statistically significant differences were found were not the same. In her case, the parametric methods found difference between both groups and time, and non-parametric between groups only. Therefore, her results may reflect more the different statistical power of the statistical methods than genuinely different results. Furthermore, Kahler argues from theory that there should be no time-effects, and therefore the non-parametric method showing no significant time effect being more accurate [147]. In this she errors in treating the "no significance" signifying "no effect", ignoring the fact that non-parametric methods would have also indicated statistical significance if the sample had been larger.

The ordinal data can be safely treated as, or actually is, interval data

A number of researchers have argued, either from real data, Monte Carlo methods, or a from the central limit theorem, that for typical data there is no need to worry about whether the data is ordinal or interval [28, 22, 133, 143]. It has also been argued that when the scores from individual Likert questions are summed, as Likert himself did, then the resulting sum will be interval data and thus appropriate for parametric methods [22, 28, 114]. Others claim that response formats with more than five categories can be considered as interval data [153].

There are several counterarguments to this position. The very argument is somewhat *ad hoc* [133] neglecting to answer the theoretical objections. Not all agree that the sum score of individual items is interval-type [140, 154, 144].

The level of measurement is not an assumption made on the parametric methods.

In mathematical statistics literature one will not find scale properties as a requirement for the use of the various statistical procedures [134, 141, 143]. Statistical procedures do make a lot of assumptions, but statisticians do not regard discrete or possibly non-uniform response ranges to be problems as long as they do not severely violate the other fundamental assumptions, such as normality [134]. Nor is the level of measurement an assumption of the parametric statistical models [141].

Not all agree with this argument, claiming that the assumptions about normality and equality of variances are something that need to be considered in addition to the level of measurement [136]. Another counterargument is that every statistical model assumes the existence of a random variable with specific distributional properties, and measurement assumptions are implicit in the definition of the random variable [146]. Violating these implicit assumptions is no different from violating the explicit assumptions.

The adherents of the representativeness philosophy have also argued that statistical methods are not a closed system, but a tool for the scientist to make scientific decisions about empirical events. Stine argued that

> ...there are operations in the statistical procedure that have no corresponding empirical event. In effect, the scientist using such a procedure would be creating structure where there is none and then using this nonsensical structure to make decisions about empirical events...A failure to recognize this role can lead to scientific decision making on the basis of nonsense [146].

Statistical methods are indifferent to how the numbers were obtained; the meaning of the numbers are outside statistics.

Statistical methods concern probability distributions of the measurement results, and are thus not deterministic. Statistical methods do not specify how these distributions were obtained; the focus is on differences and relatedness of the numbers. The meaning of numbers does not enter the picture. As early as in 1953 Lord made the point that numbers do not know where they come from [155], demonstrating this with making inferences from means and deviations from the numbers football players wore.

However, in 2010 Zand Scholten and Borsboom argued that Lord's original example was seriously flawed, in that his implicit assumption that the numbers were nominal was false, and that what was actually being analysed was bias in the setting, which is an interval property [137]. They expressed their point by saying "The numbers don't have to know where they came from; researchers have to know where they came from, since they assigned them in the first place." This however, rather misses the point that in Lord's example, the numbers were assigned first and the question on the use of the numbers rose afterwards.

This point has been argued against be asserting that the randomness of the quantities involved do not excuse them from the implications of the measurement theory, and that statistical analysis is not interesting without reference to the empirical system [136].

Parametric methods have been widely used to analyse ordinal data for decades with fruitful results. Thus the practice of doing so is supported empirically.

Yet another line of argument is the pragmatic observation that the practice of using parametric methods with ordinal data is widespread, and does produce usable and reasonable results [28, 75, 133, 114, 141, 149].

There is substantial literature suggesting that parametric statistics are appropriate [114], and parametric statistical methods tend to agree with their non-parametric counterparts [137, 141]. Thus, using parametric statistical methods with ordinal data is unlikely to be *wrong*, in that the findings or conclusions would be unreliable. Stevens himself noted this:

> As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales, for these statistics imply a knowledge of something more than the relative rank-order of data. On the other hand, for this "illegal" statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances it leads to fruitful results [132].

This argument may also be criticized by its failure to address the theoretical objections, and of being an *ad hoc* explanation for justifying the criticized but widespread practice.

5.6 Conclusions and recommendations

This chapter has reviewed some of the controversies in statistical analysis of questionnaire data (NHST, multiple comparisons, statistical analysis of ordinal data). A dissertation would be expected to find answers. However, the aim of the discussion above was not to offer any solution, but to understand what the controversies are about and why they have persisted for decades with no sign of being settled. This may be surprising, but as seen from the discussion, the nature of the controversies is such that an universally accepted solution is not likely. The issues involve different philosophical approaches to such fundamental questions as "what is the nature of measurement", and therefore the sides of the controversies do not agree on what type of argument can resolve the question. Therefore, the discussion is often a repeating of same arguments. The controversies involved are thus unresolvable, unless one assumes that others will adopt the philosophical approach one has himself -and this would be rather conceited. Many others have reached the same conclusions, that the controversy will continue. The following quotes from three different decades illustrates this point:

Michell, writing in 1986:

> "The two sides are as unrepentant as ever and show no sign of being able to appreciate the opposing point of view [145]."

Wellman and Wilkinson, in 1993:

> "We do not presume to settle a debate which has resumed for almost a half century [133]."

Norman, in 2010:

> "These findings are consistent with empirical literature dating back nearly 80 years. The controversy can cease (but likely won't) [114]."

5.6.1 Recommendations for researchers

Statistical methods are a powerful tool, which must be applied intelligently. In all three controversies reviewed above (NHST, Multiple comparison, Statistical analysis of ordinal data), most authors pointed out that much of the problems is caused by mechanistic application of statistics. Different questions are asked from the data, and these questions are appropriately answered by different methods [116].

A common theme expressed in the literature is a dismay that instructions intended as guides and advice are treated dogmatically as rules or laws that must be blindly followed [119, 111]. Helpful suggestions are treated by others as mandatory rules, possibly forcing researchers to use methods they think inappropriate or un-optimal [103]. A well-intentioned desire to keep matters simple for the students often result statistical methodology being presented as if it was, as Gigerenzer put it, "only a single hammer, rather than a toolbox [88].

The choice of statistical methods to be used needs to be made when the experiment is planned, and it depends on what is the question the study aims to find an aswer for. There are usually many ways to do right –or wrong– in statistics, so the question usually is not "which must I use" or "which is allowed", but rather "which is the best for the situation". The methods chosen will influence the necessary sample size, and in converse, the limits of available sample size may be known beforehand and they can thus affect the choice of statistical methods.

NHST alone is not sufficient for analysis, but can be useful to weed chance findings from true ones. The interesting results are often the estimates of various parameters and their uncertainty, so giving a greater emphasis of their presentation will make a paper more interesting.

One should be specific when writing. It is recommended to include the world "statistical" when discussing statistical significance of the results. When works of others is being read, care should be taken to notice when significance refers to the statistical world and when to the real world.

The question of whether the arguably ordinal data from questionnaires *can* be analysed with parametric methods seems to be settled at least empirically with a "yes." The common practice is unlikely to lead to wrong conclusions [27, 28, 75, 114, 129], even it may be questionable from the strictly theoretical point of view. Great care must be taken when any inferences made on the numbers. The question is therefore not a one of selecting between incorrect or correct methods, but of selecting the optimal one.

The controversies discussed above are very unlikely to be resolved, due to their resulting in part from different philosophical outlook. If encountering a reviewer who objects to parts of the paper on these grounds, it may be best to acquiesce. Engaging in the discussion may be intellectually stimulating, and increase the understanding of the methods involved, but it is not likely to be really fruitful. Most of the helpful guideline articles written by non-specialists, and most of the textbooks used as their sources, over simplify the issue, so one should really think before sending a simplistic articles of "here is how to do it" or "everybody is doing it wrong" type. Those who do participate in the discussion should take care to remain civil and try to understand the opposing viewpoint; far too many of the papers on the subject seems to attack the other side, or be so condescending or arrogant (see [28], for example) that they are unlikely to convince somebody who has so far been on the other side of the argument.

5.6.2 Recommendations for reviewers and editors

Reviewers and editors often see statistics as a rich area for finding mistakes, which is not surprising because statistical errors are common. However, mistaken criticism is also common. For example, Bachetti observed that

> My impression as a collaborating and consulting statistician is that spurious criticism of sound statistics is increasingly common, mainly from subject

> matter reviewers with limited statistical knowledge. Of the subject matter manuscript reviews I see that raise statistical issues, perhaps half include a mistaken criticism. [111]

One possible reason for this is the above noted tendency to see simplified guidelines as absolute rules. The reviewer and editors often have major influence on what is published and in what way, and the risk of mechanical application of rules is present here, too. In general, a reviewer should criticise statistical flaws only when he can explain how the flaws specifically detract from the study [111]. Significance testing can be useful signal of mainly the sample size of the experiment, but it is not alone sufficient for analysis, so editors and reviewers should insist on further analysis, including the size of effect estimation.

Parametric statistical methods tend to be robust; non-normality is usually not a real concern if the sample size is reasonable. Insisting using non-parametric methods because of non-normal looking plot of data is usually not warranted, although it may be a good idea to use non-parametric in addition to parametric methods used. Using parametric methods to analyse ordinal data is unlikely to lead to wrong conclusions, and should not be considered a sufficient reason to reject a paper or recommend a correction.

Sample size is important consideration, and should be considered especially if the finding seems to be that no effect exists. If the sample size is too small, there may be more serious problems than appropriateness of statistical methods.

Correction for multiple comparisons should be viewed with consideration on how the data is used. It is by no means mandatory in every case, and will increase the number of Type II errors if insisted on.

6. An analysis of lighting preference and acceptance studies

6.1 Introduction

Fotios and Houser conducted a critical analysis of 21 studies on the effect of spectral power distribution on the perception on brightness, focusing on studies that used category rating as the principal experimental methodology [8]. Their first impression was that there is no clear conclusion available. To address this, they evaluated the quality of the studies based on the quality of the research process and documentation. Their criteria for a reliable study were thorough explanation of the independent and dependent variables, complete reporting of the experimental methodologies and the collected data, and proper use of statistical analyses in forming the conclusions (or sufficient data to permit application of statistical analysis). They noted that the most common reason of being judged unreliable was incomplete reporting. All in all, they found that eleven of the twenty-one studies (i.e. 52%) were what they considered of dubious value because the published work does not present sufficient information to describe the methodology or the findings. Typical flaws were that often only the mean value of a variable is reported, and either no statistical analysis is presented, or it is inadequately described. [8]

To asses the prevailing practice of subjective lighting research, 20 studies were analysed ([32, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174]). The studies were selected from journal articles, mostly from the *Lighting Research & Technology*, which is among the most distinguished journals purely on the lighting field, and among the three journals endorsed by the CIE as the world organisation in the field of light and lighting [175]. Thus, these studies are likely thus to represent the best quality of the lighting research. As the interest is on the current or relatively recent research practice, the studies selected were published in 2005 or later. Many studies included both measured and subjectively estimated parts; only the subjective estimations of the studies are considered here. Other aspects such as performance studies are omitted in the following discussion. An overview of the studies considered is presented in tables 6.1. and 6.2.

6.2 Methods used in the studies

Twelve of the twenty studies were conducted in laboratory conditions, and involved comparisons between lighting compartments or boxes, five were field studies, and three were full-scale laboratory experiments in an environment simulating a real setting.

Most studies used a questionnaire in some form to obtain the subjective data. Other methods used rankings, Many of the studies used several types of questionnaires and response formats. Eight of the twenty studies included a paired comparison or dichotomous choice, and thirteen involved a rating task (estimation, semantic differential or Likert). These response formats ranged from five to 11 options, with a hundred-point format being used in one study [163], and one [32] involved a visual analogue scale. The most used number of response options was five (used in seven studies, followed by seven and eleven options (four and three studies). The neutral response was present in almost all cases involving more than two options.

A minority of the studies involved estimations on the reliability of the study. The methods involved null-condition tests [164, 167, 168], estimations of reproducibility and within-subject variability [177] repeated-measures subset to asses reliability [163], and estimation of resolution [164]. None of the studies were reported to be double-blind. While most used layman subjects (i.e. those not in the lighting field), some used students or the research personnel of the institution where the research was made.

6.3 Sample size and power

The number of subjects, and usually their distribution by sex and age, was usually clearly reported. The actual sample size for each comparison

Table 6.1. Overview of the studies.

Title	Field/ Laboratory	Questionnaire format	Subjects
Wilhelm et al.: Increased illuminance at the workplace: Does it have advantages for daytime shifts? [32]	Field (car factory)	Visual analog scale	23
Boissard & Fontoynont: Optimization of LED-based Light Blendings for Object Presentation [156]	Laboratory (three compartments)	Five labelled points; Determination of “the Bbest”	46
Pousset & Razet: Visual experiment on LED lighting quality with color quality scale colored samples [157]	Laboratory (two compartments)	Paired choice	35
Sándor & Schanda: Visual colour rendering based on colour difference evaluations [158]	Laboratory (two compartments)	Rating colour difference	10
Newsham et al.: in a daylit space [176]	Laboratory (full-scale office)	Rating with 7pt sematic differntial, opent questions	40
Szabó et al.: A comparative study of new solid state light sources [160]	Laboratory (two compartments)	Rating harmony and preference	9
Szabó et al.: A colour harmony rendering index based on predictions of colour harmony impression [161]	Laboratory (two compartments)	Rating harmony and preference	9
Vienot et al.: Color appearance under LED illumination: The visual judgment of observers [162]	Laboratory (compartments)	Rating colorfulness	20
Boyce et al.: Lighting quality and office work: two field simulation experiments [163]	Laboratory (full-scale office)	Many instruments ranging from yes-no to 100-point semantic differential	181

Table 6.2. Overview of the studies (continued).

Title	Field/ Laboratory	Questionnaire format	Subjects
Bullough et al.:Effects of flicker characteristics from solid-state lighting on detection, acceptability and comfort [164]	Laboratory(full scale work-station)	Yes/No and 5-point semantic differential	10
Carter & Marwaee: User attitudes toward tubular daylight guidance systems [165]	Field (office buildings)	Yes/No, and 5-point semantic differential	168
De Kort & Smolders: Effects of dynamic lighting on office workers: First results of a field study with monthly alternating settings [166]	Field (office building)	Many instruments	314
Fotios & Cheal: The effect of a stimulus frequency bias in side-by-side brightness ranking tests [167]	Laboratory (two compart -ments)	Left-Right choice on brightness	84
Fotios & Cheal: Brightness matching with visual fields of different types [168]	Laboratory (two compart -ments)	Brightness matching	10
Iszo et al.: Psychophysiological, performance and subjective correlates of different lighting conditions [169]	Laboratory (full-scale living room)	Six-point semantic differntial	30
Jaen et al.: A simple visual task to assess flicker effects on visual performance [170]	Laboratory (four boxes)	Side-by-side comparison between two boxes	10
Juslén et al.: Preferred task-lighting levels in an industrial work area without daylight [171]	Field (fuminaire factory)	5-pt Likert	25
Knight: Field surveys of the effect of lamp spectrum on the perception of safety and comfort at night [172]	Field (Residents in three locations)	5-pt semantic differential	30 – 60
Smét et al.; Optimization of colour quality of LED lighting with reference to memory colours [173]	Laboratory (objects in compart-ment)	11-pt rating	18

was often not stated; it was often implied, but not explicitly stated, that every subject responded to every item. The total number of subjects varied from 9 to 181 within one study. The mean number of subjects was 56.37, and the median 35, with the 25% quartile being 19 and 75% quartile 84 subjects. It should be noted that with a single exception [163], the studies with more than 100 subjects were field studies where the occupants of a building completed a questionnaire, thus not requiring the respondents to spend much time outside their own workplace.

The actual sample size per group depended on the experiment design. In many cases all the subjects participated in all conditions, but in others the subjects were assigned into groups and the groups compared with each other. Thus, although the total number of subjects was in one study as high as 181 [163], the group size in that study varied between 16 and 27, a much more modest figure for statistical comparison purposes. Approximately half of the studies used a sample size of less than 35 per group, which would be required to have a reasobable statistical power for medium-large effect. Only two [163, 167] of the studies involving a laboratory experiment reported a sample size larger than 50 per group. Thus, if the purpose of the study is to determine if there is a difference between the groups, using ANOVA or t-test (which was not the case in all studies considered), a typical experiment could be expected to find at most a large effect.

None of the studies presented power analysis of any kind, except one which included a sample size calculation [32].

6.4 Presentation of the results

As the subject being investigated varied greatly, no overall conclusion may be reached from these studies. This is because the criteria used to choose the studies was the methods used, not the subject or aim of research.

The results of the studies were generally reported with enough detail, typically with a mean value and the sample standard deviation. However, in number of cases only the mean was reported [158]. In other cases, the central tendency was reported in numbers, but the deviation data was presented only graphically [157, 172], and in some of the studies both the central tendency and the variation data were presented only graphically [161, 164, 178]. This is deplorable, as subsequent analysis by others is impossible without the numerical data. Therefore, it is strongly suggested that both the central tendency and variance should be presented in numerical format when an article is published.

6.5 Statistical methods used

Sixteen of the twenty studies (80 %) considered at least mentioned statistical analysis having been used. This is somewhat lower than the 94 % of articles on psychology published by APA [79], but statistical analysis is still clearly a very widely used tool in lighting research. The most commonly used statistical procedures were $t-$tests or its non-parametric equivalents, ANOVA, and Pearson's or Spearman's correlations, but other techniques mentioned included factorial analysis, chi-squared, Thurstone case V, the Cochran Q-test, linear mixed models, and Dunn-Rankin variance stable rank sums. The reasons for choosing a particular statistical method was usually not presented.

Procedures to compensate for multiple comparisons were not used, even though all of the studies involved many comparisons. The questionnaire data was usually analysed with parametric methods, and no study mentioned that they would be unsuitable for the task.

Statistical methods are usually described reasonably well, although it is rare to see the actual values of the test statistics being reported. In four cases the statistical method was described too vaguely. This is not to say that the statistical analysis was conducted incorrectly, but that it was described so incompletely that it is impossible to even asses whether the method was appropriate. For example, statistical analysis was mentioned, or p-values presented, but the actual statistical method used was not named [158, 177], or statistical analysis was done for the physiological and performance data only but apparently not for questionnaires [169, 171]. This leaves 12 out of 20 or 60 % with statistical method both used and at least adequately described.

6.6 Conclusions

The results are in line with those of Fotios and Houser discussed above [8]: slightly less than half of the studies analysed have some problems, mostly in presenting their results and inadequate description of statistical analysis. Sample sizes are often quite low for the task. Low power, however, is very common in many other fields as well [126]. It should be noted that this criticism is not meant to suggest that the methods or results were faulty; what is lacking is precision on reporting the methods, analysis and results.

7. Two experiments on survey design

7.1 Introduction

The form of a questionnaire survey can affect the results [21, 24]. The rating scale seen by the subjects can effect the results [42] and the order of the response categories is also suspected to have an effect [24], but while such effect is sometimes acknowledged in lighting literature [9], it is unclear what the effect is in the lighting context. Two experiments were conducted to see whether the order or manner of presenting the choices in a lighting survey questionnaire affects the distribution of the responses, and to estimate the magnitude of such influence if any.

These experiments and their analysis were the first actions conducted during the thesis work. The aim was also to establish a baseline; a sample of basic experiments planned as "usual", which was then analysed later again in view of the insight gained during the literature reviews. Therefore, it should not be considered as an example of how such experiment should be done. It must also be emphasized that the procedure was used to find differences in data produced by the different questionnaire versions, not to actually conduct an investigation on the actual lighting conditions. If the latter had been the object the two experiments would have been conducted differently.

7.2 Method

The test hypothesis was: Providing different response formats for the same questions will produce systematic biases or differences in the data. It was further hypothesised that these biases or differences can be either in the central tendency (mean, mean or mode), variability (standard deviation, range) or shape (skewness, kurtosis).

Two experiments were conducted with students as subjects (in hindsight, using only university student as subjects makes it arguable that the sample is not really representative of the population in general). The general procedure of the experiments were similar. In both experiments the subjects entered a room where the lighting was already on, and were given a brief introduction after which they filled a survey questionnaire. The subjects entered the room at groups of 2–4, but they were instructed to fill the forms independently, i.e. without discussing or showing the results to each other before finishing. Each subject completed the task only once, thus there is only one response per subject in the data.

The sample size was determined mostly by practical considerations, but was judged from experience to be "about the usual." (in hindsight, the sample size was too small for the task as will be seen in the discussion on statistical power below).

The same set of seven questions were used in both experiments, but the manner of presenting the response choices was varied so that two different versions were used in each experiment. All the students that came at a particular time got the same version of the questionnaire to keep them unaware of the true purpose of the experiment The questions were asked in Finnish. Translated into English, they were

1. Is there enough light in the room?

2. Is the lighting of the room suitably uniform?

3. How natural the colours in the room look?

4. I think that the colour of the lighting in the room is [the choices ranged from too warm to too cold]

5. Is there glare in the room?

6. Is the lighting in the room stimulating? [the range was from "clearly depressing" to "clearly stimulating"]

7. What is your general opinion of the lighting in the room?

A choice of five response options were given for each questions, except for question number five which used only three options (no glare, slight glare, much glare) due to fact that there cannot be too little glare. The response options followed a semantic differential format. Examples of the response formats being used are given in Figures 7.1 and 7.2.

Questions and conditions in experiment one

In the first experiment, the aim was to find out the effect of reversed order of the response options on the data. The response options followed a semantic differential format, with each response option verbally labelled. The response options were presented with an alphabet followed by a written sentence describing the response. The categories were presented vertically on top of each other so that each category formed its own line. There were two versions of the questionnaire. In the version A the responses were presented so that "clearly too little/bad/worse" was first (i.e. choice a)) and "clearly too much" was last. In the version B the order was reversed. A translation of a sample question with both categories is show in Fig. 7.1.

The room used in the experiment was an ordinary furnished two-person office room lit with tubular fluorescent lamps. The average illuminance was 225 lx (Maximum illuminance was approximately 300 lx and the minimum 100 lx). The blinds were closed during the filling of the questionnaires. Glare was estimated by measuring the UGR with the Photolux luminance mapping system. The UGR values were generally low, below 13, so it can be said that there was no significant glare in the room (the Standard EN-12464-1 (2011) specifies that the UGR to be less than 19 for most office tasks, and the measured values were well below).

A total of 50 students filled in the questionnaire. 27 used version A of the questionnaire and 23 version B. Due to varying group sizes completely equal number of versions could not be enforced.

Questions and conditions in experiment two

The aim of the second experiment was to find out how the different format of response options affects the results. In the second experiment the response categories were presented in two alternate forms. The version C was identical with the version A used in the experiment one, but as the experiments were conducted in different room with different lighting conditions the results cannot be combined with those of the first experiment. In the other (the version D) just the endpoints of the semantic differen-

VERSION A:

7. What is your general opinion of the lighting in the room?

a) Lighting in the room is clearly worse than in working places usually.

b) Lighting in the room is slightly worse than in working places usually.

c) Lighting in the room is as good as in working places usually.

d) Lighting in the room is slightly better than in working places usually.

e) Lighting in the room is clearly better than in working places usually.

VERSION B:

7. What is your general opinion of the lighting in the room?

a) Lighting in the room is clearly better than in working places usually.

b) Lighting in the room is slightly better than in working places usually.

c) Lighting in the room is as good as in working places usually.

d) Lighting in the room is slightly worse than in working places usually.

e) Lighting in the room is clearly worse than in working places usually.

Figure 7.1. Sample questions in experiment one.

VERSION C:

7. What is your general opinion of the lighting in the room?

a) Lighting in the room is clearly worse than in working places usually.

b) Lighting in the room is slightly worse than in working places usually.

c) Lighting in the room is as good as in working places usually.

d) Lighting in the room is slightly better than in working places usually.

e) Lighting in the room is clearly better than in working places usually.

VERSION D:

7. What is your general opinion of the lighting in the room, compared with the usual lighting in working places?

clearly worse ○ ○ ○ ○ ○ clearly better

Figure 7.2. Sample question in experiment two.

tial range were labelled with text and the responses were presented as marks between the labels. The order of the options was the same in both versions. An example of the versions is shown in Fig. 7.2.

The room in the experiment two was not the same as in the experiment one, but similar a furnished two-person office room with fluorescent lighting. However, because the experiments were not conducted in the same room, the results are not directly comparable. The average illuminance of the room used in the experiment two was 400 lx, and some daylight was allowed in this room because of broken Venetian blinds in the upper part of the window. UGR was not measured, since the variable daylight would have made such measures time-dependent, but direct sunlight was not present during the experiments.

A total of 32 subjects returned the forms. 13 filled the version C and 19 the version D. Due to varying group sizes completely equal number of versions could not be enforced.

7.3 Results

All the results were transformed into numerical range from 1 to 5 so that the response options with the same content received the same number (e.g. for question 7, “clearly worse” was coded as 1, and “clearly better” as 5 for all versions). The only exception was the question five, which had

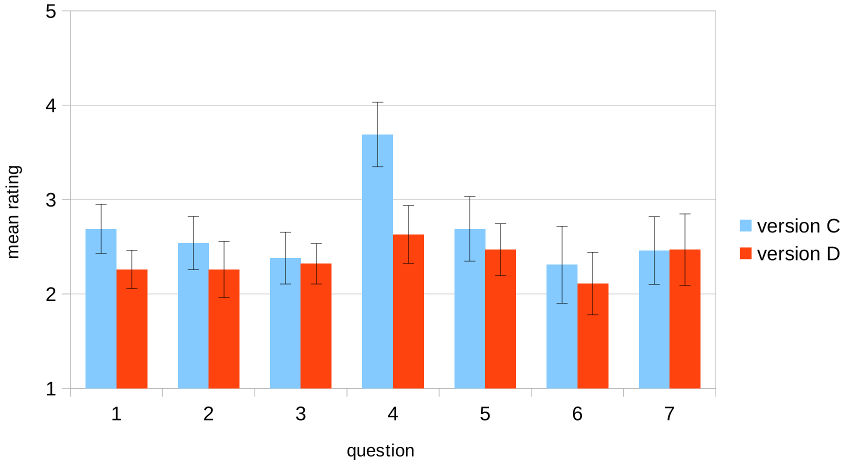


Figure 7.6. The mean ratings and their 95 % confidence intervals of the second experiment

Table 7.2. Results for the experiment 2 recoded into the common scale.

	Version C (N=13)			Version D (N=19)		
	mean	median	mode	mean	median	mode
Q1: Is there enough light in the room?	2.69	3	3	2.26	2	2
Q2: Is the lighting of the room suitably uniform?	2.54	3	3	2.26	2	2
Q3: How natural the colours in the room look?	2.38	2	2	2.32	2	2
Q4: I think that the colour of the lighting in the room is [too warm...too cold]	3.69	4	4	2.63	3	3
Q5: Is there glare in the room?	2.69	3	3	2.47	3	3
Q6: Is the lighting in the room stimulating?	2.31	2	2	2.11	2	2
Q7: What is your general opinion of the lighting in the room?	2.46	2	2	2.47	2	2

Table 7.3. Differences in central tendency in the Experiment 1.

	Version A – Version B		
	mean	median	mode
Q1: Is there enough light in the room?	0.114	+1	+1
Q2: Is the lighting of the room suitably uniform?	0.090	0	0
Q3: How natural the colours in the room look?	-0.029	0	–
Q4: I think that the colour of the lighting in the room is [too warm...too cold]	0.129	0	-1
Q5: Is there glare in the room?	0.300	0	0
Q6: Is the lighting in the room stimulating?	-0.483	-1	-2
Q7: What is your general opinion of the lighting in the room?	-0.264	-1	-1

systematic effect across the questions in the first experiment; the mean of the mean differences is -0.020. Four questions produced higher mean with version A and three with version B of the questionnaire, which is as close to even split as it is possible to get with seven questions without the differences of mean being zero for any questions. A central tendency expressed with median or mode is harder to judge, because a single change of one step by a single responder can change the median or mode by a whole point. In the first experiment, the version A of the questionnaire produced higher median in one questions out of seven but lower in two. The Modes were higher in one but lower in three questions.

In the second experiment the absolute value of the difference of the means ranged from 0.012 to 1.061, the mean being 0.324. Here there may be a slight systematic effect; the mean of the mean differences is 0.320, and in only one question (Question 7: “General opinion of the lighting”) the version D had higher mean than the version C. In three questions the version C produced higher median and mode than the version D and never a lower value. However, the large difference in question four may influence the mean; if question four is omitted from the data the mean of the differences is just 0.197, which would still suggest the possibility of a small systematic shift.

Table 7.4. Differences in central tendencies in the Experiment 2

	Version C – Version D		
	mean	median	mode
Q1: Is there enough light in the room?	0.429	+1	+1
Q2: Is the lighting of the room suitably uniform?	0.275	+1	+1
Q3: How natural the colours in the room look?	0.069	0	0
Q4: I think that the colour of the lighting in the room is [too warm...too cold]	1.061	+1	+1
Q5: Is there glare in the room?	0.219	0	0
Q6: Is the lighting in the room stimulating?	0.202	0	0
Q7: What is your general opinion of the lighting in the room?	-0.012	0	0

Statistical analysis was done with SPSS version 20 software. Means were compared with the independent samples Student's t-test, because it is the simplest and most used method for comparing means of two groups. For comparison, two non-parametric analogues for the t-tests were used: the Mann-Whitney U-test (MW) and Kolmogorov-Smirnov (KS) test. Two-tailed tests were used in each case. The results are shown in Table 7.5 (it should be noted that in a general research article, significance is better presented with some way taking less space. Here it is necessary because the statistical methods are compared, which is usually not the case in a research article) . No adjustments for multiple comparisons were made. This is justified as the aim was to find consistent patterns emerging from the results, not individual differences. Had the Bonferroni correction been used, only the Question four in the second experiment would have had statistically significant differences.

The t-test detected no statistically significant differences in means between the versions in the first experiment. The MW test compares whether the rankings of the distributions are different, i.e. whether the values belonging to one group of the samples tend to be larger than those belonging to the other. In the experiment one the MW test reports statistically significant ($p \leq 0.050$) difference only for Question five ("Is there glare in the

Table 7.5. Statistical significance (p-values) of differences between questionnaire versions for Mann-Whitney U test (MW), the Kolmogorov-Smirnov test (KS) and the Student's t-test (t) for independent samples.

	Experiment 1			Experiment 2		
	MW	KS	t	MW	KS	t
Q1: Is there enough light in the room?	0.362	0.917	0.463	0.041	0.116	0.015
Q2: Is the lighting of the room suitably uniform?	0.562	1.000	0.532	0.209	0.602	0.215
Q3: How natural the colours in the room look?	0.821	1.000	0.878	0.762	1.000	0.699
Q4: I think that the colour of the lighting in the room is [too warm...too cold]	0.640	0.779	0.522	<0.001	0.001	<0.001
Q5: Is there glare in the room?	0.036	0.236	0.064	0.305	0.753	0.335
Q6: Is the lighting in the room stimulating?	0.050	0.210	0.085	0.495	1.000	0.455
Q7: What is your general opinion of the lighting in the room?	0.157	0.481	0.218	0.940	1.000	0.966

room, with p = 0.036 and U = 0.255), and six ("Is the lighting stimulating", p = 0.050, U = 214.5). The Kolmogorov-Smirnov test revealed no statistically significant differences between the versions. However, this does not reveal a pattern: The questionnaire version A (where the responses options were ordered from "clearly worse to clearly better") produced higher mean score than the version B (with the reverse order of response options) for Question five, but this was reversed for the Question six.

In the second experiment the MW test reported statistically significant ($p \leq 0.050$) differences for questions one ("Is there enough light in the room", p = 0.041, U =71.5) and four ("I think that the colour of the lighting in the room is [cold/warm]", p < 0.001, U = 33). The Kolmogorov-Smirnov test reports statistically significant differences only for question four (p = 0.001, Z = 1.991). The t-test reported statistical significance for questions one (p = 0.018, t = 2.572) and four p < 0.001, t = 0.444). The questionnaire version C (all response options labelled) produced higher answers with higher means than the version D (only end-points are labelled) for both Questions one and four, which suggests a pattern; the labelled version produces data with higher means.

Confidence intervals have been proposed as an alternative for statistical null hypothesis significance testing. The 95 % confidence intervals are presented in the Figures 7.4 and 7.6. Cumming and Finch suggest that when estimating independent means, as here, graphically with 95 % confidence intervals, the $p-$value of about 0.05 corresponds to approximately overlap of no more than about half of the marigin of error, and $p \leq 0.01$ when the overlap is nearly zero, or ther is a visible gap[109]. A quick graphical comparison shows that the 95 % confidence intervals always overlap in the results of the first experiment, and for none is the overlap clearly less than half the marigin of error. (Question six quite clear, though). In the second experiment, only the 95 % CIs for question four are clearly not overlapping, and the question one is close to the approximately zero overlap. Thus, on basis of of analysing the CIs by eye, one would expect that the p-value is definitely below 0.01 for question four and close but more than 0.01 in question one in the second experiment. Such certainty is not achieved in the first experiment.

Effect size and statistical power

There are several indices to estimate effect size. When the analysis is performed with the t-test, the recommended method for effect size is Cohen's d [92], which is calculated with

$$d = \frac{\Delta\mu}{s}, \tag{7.1}$$

where

$\Delta\mu$ is the difference between the means[1] and

s is the pooled standard deviation, given by

$$s = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}, \tag{7.2}$$

where

N_i is the sample size of group i and

s_i is the sample standard deviation of the sample i.

Cohen himself advised the researcher to use his experience in judging which values of d should be considered to represent a small or large effect

[1] Here the absolute value i.e. the magnitude of the difference was used. In some cases the direction of the difference and hence the sign of the effect may be of interest.

Table 7.6. Effect sizes expressed as Cohen's d's.

	Experiment 1	Experiment 2
Q1: Is there enough light in the room?	0.21	0.93
Q2: Is the lighting of the room suitably uniform?	0.18	0.46
Q3: How natural the colours in the room look?	0.05	0.14
Q4: I think that the colour of the lighting in the room is [too warm...too cold]	0.18	1.60
Q5: Is there glare in the room?	0.55	0.35
Q6: Is the lighting in the room stimulating?	0.50	0.27
Q7: What is your general opinion of the lighting in the room?	0.36	0.02

in the particular context, but the usual convention is to use values of 0.2 for a small effect, 0.5 for a medium and 0.8 for large effect (this is another example of the hard–soft paradox mentioned in the beginning of the Chapter 5).

The effect sizes for both experiments are shown in the Table 7.6. In the first experiment the effect size ranged from 0.05 to 0.55, that is from insignificant to a medium effect size, if judged by the usual convention. The mean effect size was 0.29 and the median 0.21. Thus, the effect, if any, in the first experiment is in the what is usually considered small-medium region. In the experiment two the effect size ranged from 0.02 to 1.60 with the mean being 0.54 and the median 0.35. In two questions (one and four) the effect was clearly a large one, but in the other questions the effect size was from none to small-medium by usual conventions. The cases where statistical tests showed statistical significance (question five in the first experiment and questions one and four in the second) the effect size was medium or large. It should be recalled that the observed effect size is only an estimate of the true one and has its own uncertainty.

Statistical power is the probability of rejecting the Null hypothesis when it is in fact false. The *a priori* power for both experiments was calculated for two-tailed t-test with G*Power 3.1.4 software [117] , assuming a large (d = 0.8), medium (d = 0.5) and small ($d = 0.2$) effect size and the actually used sample sizes.

The calculated power in the first experiment (sample size of of 50 with groups of 27 and 23) was 0.79 for large, 0.41 for medium and 0.11 for a small effect. For the second experiment (sample size of 32 with groups of 13 and 19) the corresponding power was 0.58 for large, 0.27 for medium and 0.08 for a small effect. Thus, the sample size was clearly smaller than the required to achieve the power of 0.8 even for a large effect. Achieving statistical power of 0.8 would have required a total sample size of at least 52 for large, 128 for medium and 788 for a small effect (assuming equal-sized sample groups). Thus, the statistical power was clearly inadequate even for a medium effect – and most of the observed effect sizes were medium or small. The sample sizes were less than the 30 per group recommended in the literature [120], but still larger than that in many of studies published in the leading journal in the lighting field (see chapter 6).

Standard deviation

The observed standard deviations according to the question version are presented in Table 7.7. In the experiment one the absolute value of the difference in standard deviation ranged from 0.001 to 0.221 with mean of 0.066 and median of 0.044. The mean of the differences was -0.048 and the median -0.001. This indicates that reversing the response option order did not produce a systematic effect on standard deviations in the experiment one.

In the second experiment the absolute value of the difference of the observed standard deviations ranged from 0.014 to 0.181 with the mean of 0.065 and median 0.028. The mean of the differences was -0.040 and the median 0.014. These values are very similar to those in the experiment one, and show no systematic effect on the standard deviation.

The variances within experiment (i.e. results for questionnaire versions A and B were compared with each other, as were C and D, but not A and C or B and D) were compared with Levene's test for equality in variances. The test detected statistically significant difference only for the question five "Is there glare in the room" in the experiment one (Levene statistics is 4.894 and p = 0.032), and no differences in the experiment two (However, it must be noted that the sample sizes are quite small). All in all, the presentation of the questionnaires did not produce any observable systematic effect on the variance.

Table 7.7. Observed standard deviation by version.

	Experiment 1			Experiment 2		
	A	B	difference (A – B)	C	D	difference (C – D)
Q1: Is there enough light in the room?	0.501	0.593	-0.092	0.480	0.452	0.028
Q2: Is the lighting of the room suitably uniform?	0.509	0.499	0.010	0.519	0.653	-0.134
Q3: How natural the colours in the room look?	0.679	0.635	0.044	0.506	0.478	0.028
Q4: I think that the colour of the lighting in the room is [too warm...too cold]	0.594	0.815	-0.221	0.630	0.684	-0.054
Q5: Is there glare in the room?	0.506	0.593	-0.087	0.630	0.612	0.018
Q6: Is the lighting in the room stimulating?	0.974	0.964	0.010	0.751	0.737	0.014
Q7: What is your general opinion of the lighting in the room?	0.735	0.736	-0.001	0.660	0.841	-0.181

Skewness and kurtosis

Looking at the graphical presentation of the data in Figures 7.3. and 7.5. suggests that the data is scattered and that skewness and kurtosis varied very much. This is indeed what happens as seen in the Tables 7.8 and 7.9.

In the first experiment the absolute value of difference of skewness varied from 0.133 to 1.697, with an average of 0.957. However, the mean of the difference of skewness was just 0.014 and median 0.133. This indicates that the two versions of the questionnaires produced no systematic effect on skew in the first experiment.

In the experiment two the absolute values of difference between skewness ranged from 0.028 to 2.116 with mean of 0.988. The mean of the difference was -0.85 and median -1.178. Furthermore, for five out of seven questions the version D produced more positive skew than version C. This suggests that version D produced data with more positive (or less negative) skew than version C. However, as noted above, there was a statistically significant difference in distributions only in questions one and four.

Kurtosis also varied greatly. In the first experiment the absolute value of the difference of kurtosis between the versions ranged from 0.109 to as much as 5.571 the mean being 1.374. This value, however, depends

Table 7.8. Skewness by version.

	Experiment 1			Experiment 2		
	A	B	difference (A–B)	C	D	difference (C–D)
Q1: Is there enough light in the room?	-0.399	0.806	-1.205	-0.946	1.170	-2.116
Q2: Is the lighting of the room suitably uniform?	0.079	0.477	-0.398	0.539	0.862	-0.323
Q3: How natural the colours in the room look?	0.530	0.340	0.19	-0.157	1.021	-1.178
Q4: I think that the colour of the lighting in the room is [too warm...too cold]	-0.122	-0.255	0.133	-2.051	-0.532	-1.519
Q5: Is there glare in the room?	-2.322	-0.625	-1.697	-2.051	-0.703	-1.348
Q6: Is the lighting in the room stimulating?	1.022	-0.578	1.6	0.784	0.756	0.028
Q7: What is your general opinion of the lighting in the room?	1.100	-0.376	1.476	1.191	0.718	0.473

greatly on the large difference on question five, and if this question is omitted the mean becomes only 0.578. The mean of the difference was 0.769 (-0.027 if question five is omitted) and median 0.109 (-0.012 if question five is omitted), so experiment one did not show any clear systematic effect on kurtosis between the versions.

The absolute values of the difference of kurtosis between questionnaire version C and D ranged from 0.265 to 4.023 with the mean being 1.420. The mean of the difference was 0.857 and the median -0.265. When the differences are examined (Table 7.9) it is clear again that the mean is heavily influenced by the two large positive values (questions four and five). In both experiments the number of questions producing positive or negative values of kurtosis was three of one and four of the other, i.e. as close to even as is possible with seven questions and no ties. Thus, these results give no justification to say that the questionnaire versions produced data with different kurtosis.

Response style effects

The used sample sizes are rather small to estimate response style effects. it can be seen from Figures 7.3 and 7.5 that responses using the extreme

Table 7.9. Kurtosis by version.

	Experiment 1			Experiment 2		
	A	B	difference (A–B)	C	D	difference (C–D)
Q1: Is there enough light in the room?	-1.994	-0.218	-1.776	-1.339	-0.718	-0.621
Q2: Is the lighting of the room suitably uniform?	-2.160	-1.951	-0.209	-2.056	-1.419	-0.637
Q3: How natural the colours in the room look?	-0.650	-0.517	-0.133	-2.364	1.915	-0.449
Q4: I think that the colour of the lighting in the room is [too warm...too cold]	-0.347	-1.432	1.085	3.711	0.593	3.118
Q5: Is there glare in the room?	5.101	-0.470	5.571	3.711	-0.312	4.023
Q6: Is the lighting in the room stimulating?	0.003	-0.106	0.109	1.223	1.488	-0.265
Q7: What is your general opinion of the lighting in the room?	-0.241	-0.975	0.734	0.645	-0.185	0.830

end of the option range were rare in all cases except for the question about glare, where “no glare” was the most chosen option in both experiments. Therefore, extreme response style did not seem to occur; a possibility of middle response remains, with the subjects tending to answer always with a qualified response option. In the second experiment, the questionnaire version with only end-points labelled seems to produce extreme responses slightly more often than the fully labelled version, but the rate of extreme responses is so small that firm conclusions cannot be made from these results.

7.5 Conclusions

The general conclusion is that the form of questionnaire can affect the results. This effect may make difficult comparing directly the results from different response formats at least quantitatively. Presenting the response options in reverse order did not suggest any systematic effect on the central tendency of the data. Although there were indications of possible small to medium-sized effects (which were not statistically significant due to small sample size), they did not show a clear pattern or one version of the questionnaire producing responses consistently biased to one direction when compared to the other.

Labelling the response options verbally produced responses with central tendency slightly more to the more/better than when only the endpoints of the response range were labelled.. Such result with data from Likert-type response format can be explained by acquiescence effects, but this explanation is not as readily applicable here where the data was obtained with a semantic differential format. It would be interesting to repeat the second experiment with the response option order reversed, to see if the effect remains.. This result is similar to those Weijters et al. obtained with Likert-type data in marketing survey context [43] It is impossible to say on the basis of these results whether the results from the labelled or unlabelled version of the questionnaire reflected more accurately the respondent's state of mind. Literature from other fields [21, 44, 43] suggests that labelling every option verbally makes it easier for the subjects to understand the options, and that the subjects prefer formats where every option is labelled.

The MW test performed better in the experiment one than the t-test in finding statistically significant differences. In the experiment two the two performed identically. The KS test was much behind the MW and t-tests in both experiments. The often-cited better statistical power of the t-test does not apply when the distribution of data is much from the normal, and the MW test may be in fact much more powerful in these cases [128, 92, 129, 27]. This effect was not observed in the present experiments. Using the parametric t-test to analyse the questionnaire data led to the same conclusions as its non-parametric equivalent, the MW-test. Therefore, these results do not support the view that parametric methods are inappropriate for analysing questionnaire data. In conclusion, neither MW nor t-test was observed to have a clear advantage in power or validity. This result is in agreement with those from literature from other fields (e.g. [114, 130, 27]).

Using the 95 % Confidence intervals to find differences is easy and intuitive to do, but is less powerful than the MW or t-tests, as suggested in the literature [77].

Neither of the experiments demonstrated any systematic effect on variance, skewness or kurtosis of the resulting data. Thus these experiments give no basis to say that some form of the questionnaire produces better quality data. Literature supports using the labelled version as it is known that reliability and validity of a questionnaire can be improved if all points on the scale are labelled with words [21, 24]. As each subject participated only once, the effect of response format on test-retest reliability could not be studied.

The analysis of these experiments underlined the need to estimate the necessary sample during the planning of the experiments. Effects from the response format are likely to be subtle, thus investigating them requires a large sample if good statistical power is desired.

8. Conclusions and recommendations

This work has discussed many of the factors the researcher has to consider when designing subjective lighting experiments. Lighting preference and acceptance are attitudes, and thus the research methods are similar to ones measuring attitudes in other fields. The literature on investigating attitudes in other fields of study (e.g. psychology, social sciences, marketing research) is rich and contain excellent research methodology papers as well. There is also a small number of papers on specifically subjective lighting research methodology.

The present work did not find a general answer to the inspirational question "How should subjective lighting evaluation research be desinged and conducted." It is not possible to give a general rule on how one must design an acceptance or preference study or analyse the data. It is not even desirable to do so; after all, the researcher has a large number of potential tools available, and there is no grounds to categorically limit them to one or few. Virtually all the authors in the literature reviewed stress the importance of giving enough attention to the experimental design. Decisions made in this stage will affect the available tools for the data analysis, the sample size needed, and other factors. The importance of careful planning is echoed by those writing on the statistical issues as well; much of the criticism for poor statistical practices arise from not thinking well enough what will be asked from the data once it is obtained.

There are many potential sources for biases in questionnaire research. Many of the most common are described in the chapters 3 and 4. The research instrument, setting and protocol can be designed to avoid or lessen these effects, which will improve the quality of the research. Usually this does not even require significant extra work, just paying attention to the experimental design.

The use of non-parametric statistical methods for analysing the data from questionnaires is controversial in some fields. The results obtained during this work do not support the view that it is wrong to use parametric methods for questionnaire data.

On the basis of the results of this study, attention should be paid in the following points during the research work:

1. The researcher should ask the fundamental questions – “what is the purpose of the research?”, “what is the research question to be answered?”, “what kind of answer am I after, a yes–no, ranking, rating, or something else”, “is there a critical parameter to be estimated, and if so, what is the necessary precision”, “is there already an existing instrument that could be used?”, “how are the results going to be analysed?”

2. After having answers to the fundamental questions, the instrument must be designed. The choice of response format is driven by two requirements; the response format must allow enough latitude for the subjects to express themselves adequately, and generate enough variance among the subjects to allow subsequent analysis. There are many response formats available. Likert and semantic differential types are most commonly used. The most commonly used number of response options is five to seven, and their use is justified by the literature review of this work. Omitting the neutral option may be advantageous in many cases, but needs to be carefully considered. The results of the experiments done during this work suggest that a questionnaire with every response option labelled may produce slightly different results than if only the end-points are labelled, and the literature contains a number of studies which suggest that the data characteristics, and respondent satisfaction, will improve if all the response options are labelled. However, this may reinforce acquiescence effects, especially if Likert-type response format is used.

3. Sample size and statistical power needs to be considered during the experimental design. The criterion for the sample size can be the probability of obtaining statistical significant results, or the degree of uncertainty or tolerable margin of error in parameter estimation. Often the aim is to achieve statistical significance. Even when the usual criteria of sample size calculations would indicate an impractically large sample, it should be estimated to what degree the power or accuracy is sacrificed

if a smaller sample size is selected. The research paper written about the study should include some discussion on the sample size.

4. After the experiments are designed, they should be piloted. The reliability and resolution of the experiments should be considered (and written about in the final paper), and the set-up should include the null-condition test if at all possible. It should be considered if it is possible to conduct the test as a double-blind experiment.

5. Attention should be paid on the motivation of the subjects; the importance of their honest participation should be stressed, and the concern on how they perform eased. The possibility of response style effects should be kept in mind when the questionnaire is developed and the results analysed. The present experiments showed a possible middle range response style effects.

6. Statistics are a servant and not a master. Statistics are not be applied mindlessly, but after a careful consideration on what is being done and what is the best way to do it. The statistical methods that will be used to analyse the data need to be chosen when the experiment is being designed, and the choice will affect the experiment and the necessary sample size. There are usually many ways to do right – or wrong– in statistics, so the question usually is not "which must I use" or "which is allowed", but rather "which is the best for the situation".

7. When the research results are published, care should be taken to present the result and analysis in detail. Sufficient information should be provided to allow the readers to analyse the results further. This would also help in combining the results of several studies. Statistical significance is usually less interesting that the actual values obtained. Therefore statistical significance is best communicated in some simple way not taking too much space (e.g. typographic means or special symbols) rather than tables containing nothing more than p-values. Discussion of the methodology choices made would be interesting.

The present study was centred on theoretical aspects via a literature surveys. It was also necessary to limit the task to a manageable amount of work, and thus the scope was limited to comparison of means by different groups, and not, for example, correlation models or factorial analysis. It must be admitted that the empirical part of this study is rather limited, and there remains much room for further research on lighting research methodology. One interesting question would be the question between rating and ranking; do these approaches lead to a similar conclusions about preference. A comparison between Likert and Semantic differential response formats in lighting context might also be an interesting subject to study. However, a sample size calculation is very strongly recommended before such research is conducted. These effects are likely to be subtle and therefore require a quite large sample.

References

[1] A. Finki, *How to Analyze Survey Data*. Sage Publications, 1995.

[2] J. Flynn, C. Hendrick, T. Spencer, and O. Martyniuk, “A guide to methodology procedures for measuring subjective impressions in lighting,” *Journal of the Illuminating Engineering Society*, vol. 8, no. 2, pp. 95–110, 1979.

[3] M. Rea, “Calibration of subjective scaling responses,” *Lighting Research & Technology*, vol. 14, no. 3, pp. 121–129, 1982.

[4] D. Tiller and M. Rea, “Semantic differential scaling: prospects in lighting research,” *Lighting Research & Technology*, vol. 24, no. 1, pp. 43–51, 1992.

[5] K. W. Houser and D. K. Tiller, “Measuring the subjective response to interior lighting: paired comparisons and semantic differential scaling,” *Lighting Research & Technology*, vol. 35, no. 3, pp. 183–195, 2003.

[6] S. A. Fotios, “Lamp colour properties and apparent brightness: a review,” *(Lighting Research & Technology)*, vol. 33, no. 3, pp. 163–178, 2001.

[7] ——, “An error in brightness matching associated with the application of dimming,” *Lighting Research & Technology*, vol. 33, no. 4, pp. 223–229, 2001.

[8] S. Fotios and K. Houser, “Research methods to avoid bias in categorical ratings of brightness,” *Leukos*, vol. 5, no. 3, pp. 167–181, 2009.

[9] D. Atli and S. Fotios, “Rating spatial brightness: does the number of response categories matter?” *Ingineria Iluminatului*, vol. 13, no. 1, pp. 15 – 28, 2011.

[10] A. N. Oppenheim, *Questionnaire design, Interviewing and Attitude measurement*. Continuum International Publishing Group, 2000.

[11] P. Murray, “Fundamental issues in questionnaire design,” *Accident and Emergency Nursing*, vol. 7, no. 3, pp. 148–153, 1999.

[12] D. Javeline, “Response effects in polite cultures: a test of acquiescence in Kazakhstan,” *Public Opinion Quarterly*, vol. 63, pp. 1–28, 1999.

[13] I. McDowell, *Measuring health: a guide to rating scales and questionnaires*. Oxford University Press, 2006.

[14] R. S. Hunter, *The measurement of Appearance*. Wiley-Interscience, 1975.

[15] I. Ajzen, "Nature and operation of attitudes," *Annual review of psychology*, vol. 52, no. 1, pp. 27–58, 2001.

[16] T. D. Wilson, S. Lindsey, and T. Y. Schooler, "A model of dual attitudes." *Psychological Review*, vol. 107, no. 1, p. 101, 2000.

[17] S. H. Kim and H. Tokura, "Effects of menstrual cycle and room temperature on color preference," *Biological Rhythm Research*, vol. 29, no. 2, pp. 151–158, 1998.

[18] H. Sook Kim and T. Hiromi, "Cloth color preference under the influence of face cooling," *Journal of Thermal Biology*, vol. 23, no. 6, pp. 335–340, 1998.

[19] S. H. Kim and H. Tokura, "Visual alliesthesia – cloth color preference in the evening under the influence of different light intensities during the daytime," *Physiology & Behavior*, vol. 65, no. 2, pp. 367–370, 1998.

[20] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: a critical review of the literature and recommended remedies." *Journal of Applied Psychology*, vol. 88, no. 5, p. 879, 2003.

[21] J. A. Krosnick, "Survey research," *Annual Review of Psychology*, vol. 50, no. 1, pp. 537–567, 1999.

[22] J. Carifio and R. J. Perla, "Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes." *Journal of Social Sciences*, vol. 3, no. 3, pp. 106–116, 2007.

[23] N. C. Schaeffer and S. Presser, "The science of asking questions," *Annual Review of Sociology*, vol. 29, pp. 65–88, 2003.

[24] J. A. Krosnick, "Maximizing questionnaire quality," *Measures of Political Attitudes*, vol. 2, pp. 37–58, 1999.

[25] J. A. Krosnick and D. F. Alwin, "An evaluation of a cognitive theory of response-order effects in survey measurement," *Public Opinion Quarterly*, vol. 51, no. 2, pp. 201–219, 1987.

[26] J. Dawes, "Do data characteristics change according to the number of scale points used?" *International Journal of Market Research*, vol. 50, no. 1, pp. 61 –77, 2008.

[27] J. C. de Winter and D. Dodou, "Five-point likert items: t-test versus Mann-Whitney-Wilcoxon," *Practical Assessment, Research & Evaluation*, vol. 15, no. 11, pp. 1–12, 2010.

[28] J. Carifio and R. Perla, "Resolving the 50-year debate around using and misusing Likert scales," *Medical Education*, vol. 42, no. 12, pp. 1150–1152, 2008.

[29] S. Jamieson, "Likert scales: how to (ab)use them," *Medical Education*, vol. 38, no. 12, pp. 1217–1218, 2004.

[30] C. Laurentin, V. Bermtto, and M. Fontoynont, "Effect of thermal conditions and light source type on visual comfort appraisal," *Lighting Research & Technology*, vol. 32, no. 4, pp. 223–233, 2000.

[31] D. Loe, K. Mansfield, and E. Rowlands, “A step in quantifying the appearance of a lit scene,” *Lighting Research &Technology*, vol. 32, no. 4, pp. 213–222, 2000.

[32] B. Wilhelm, P. Weckerle, W. Durst, C. Fahr, and R. Röck, “Increased illuminance at the workplace: does it have advantages for daytime shifts?” *Lighting Research & Technology*, vol. 43, pp. 185–199, 2010.

[33] D. F. Alwin and J. A. Krosnick, “The reliability of survey attitude measurement the influence of question and respondent attributes,” *Sociological Methods & Research*, vol. 20, no. 1, pp. 139–181, 1991.

[34] J. A. Krosnick, “Response strategies for coping with the cognitive demands of attitude measures in surveys,” *Applied Cognitive Psychology*, vol. 5, no. 3, pp. 213–236, 1991.

[35] C. Preston and A. Colman, “Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences,” *Acta Psychologica*, vol. 104, no. 1, pp. 1–15, 2000.

[36] K. F. Schulz, I. Chalmers, R. J. Hayes, and D. G. Altman, “Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials,” *The Journal of the American Medical Association*, vol. 273, no. 5, pp. 408–412, 1995.

[37] L. Wood, M. Egger, L. L. Gluud, K. F. Schulz, P. Jüni, D. G. Altman, C. Gluud, R. M. Martin, A. J. Wood, and J. A. Sterne, “Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study,” *British Medical Journal*, vol. 336, no. 7644, p. 601, 2008.

[38] V. Bebarta, D. Luyten, and K. Heard, “Emergency medicine animal research: does use of randomization and blinding affect the results?” *Academic Emergency Medicine*, vol. 10, no. 6, pp. 684–687, 2003.

[39] S. J. Day and D. G. Altman, “Blinding in clinical trials and other studies,” *British Medical Journal*, vol. 321, no. 7259, p. 504, 2000.

[40] T. R. Hinkin, “A brief tutorial on the development of measures for use in survey questionnaires,” *Organizational research methods*, vol. 1, no. 1, pp. 104–121, 1998.

[41] J. Hofmans, P. Theuns, S. Baekelandt, O. Mairesse, N. Schillewaert, and W. Cools, “Bias and changes in perceived intensity of verbal qualifiers effected by scale orientation,” in *Survey Research Methods*, vol. 1, no. 2, 2007, pp. 97–108.

[42] N. Schwarz, B. Knäuper, H. Hippler, E. Noelle-Neumann, and L. Clark, “Rating scales numeric values may change the meaning of scale labels,” *Public Opinion Quarterly*, vol. 55, no. 4, pp. 570–582, 1991.

[43] B. Weijters, E. Cabooter, and N. Schillewaert, “The effect of rating scale format on response styles: the number of response categories and response category labels,” *International Journal of Research in Marketing*, vol. 27, no. 3, pp. 236–247, 2010.

[44] R. Garland, “A comparison of three forms of the semantic differential,” *Marketing Bulletin*, vol. 1, no. 1, pp. 19–24, 1990.

[45] T. Johnson, P. Kulesa, Y. I. Cho, S. Shavitt *et al.*, “The relation between culture and response styles evidence from 19 countries,” *Journal of Cross-Cultural Psychology*, vol. 36, no. 2, pp. 264–277, 2005.

[46] D. K. Tiller, “Toward a deeper understanding of psychological aspects of lighting,” *Journal of the illuminating Engineering Society*, vol. 19, no. 2, pp. 59–65, 1990.

[47] B. Weijters, H. Baumgartner, and N. Schillewaert, “Reversed item bias: an integrative model.” *Psychological Methods*, vol. 18, no. 3, p. 320, 2013.

[48] G. Moors, “Exploring the effect of a middle response category on response style in attitude measurement,” *Quality & Quantity*, vol. 42, no. 6, pp. 779–794, 2008.

[49] G. A. Miller, “The magical number seven, plus or minus two: some limits on our capacity for processing information.” *Psychological Review*, vol. 63, no. 2, p. 81, 1956.

[50] F. Wuyts, M. De Bodt, and P. Van de Heyning, “Is the reliability of a visual analog scale higher than an ordinal scale? an experiment with the grbas scale for the perceptual evaluation of dysphonia,” *Journal of Voice*, vol. 13, no. 4, pp. 508–517, 1999.

[51] D. V. Cicchetti, D. Shoinralter, and P. J. Tyrer, “The effect of number of rating scale categories on levels of interrater reliability: a monte carlo investigation,” *Applied Psychological Measurement*, vol. 9, no. 1, pp. 31–36, 1985.

[52] S. M. Nowlis, B. E. Kahn, and R. Dhar, “Coping with ambivalence: The effect of removing a neutral option on consumer attitude and preference judgments,” *Journal of Consumer Research*, vol. 29, no. 3, pp. 319–334, 2002.

[53] P. E. Converse, *The Nature of Belief Systems in Mass Publics*. Survey Research Center, University of Michigan, 1962.

[54] D. F. Alwin and J. A. Krosnick, “The measurement of values in surveys: a comparison of ratings and rankings,” *Public Opinion Quarterly*, vol. 49, no. 4, pp. 535–552, 1985.

[55] J. A. Krosnick, A. L. Holbrook, M. K. Berent, R. T. Carson, W. M. Hanemann, R. J. Kopp, R. C. Mitchell, S. Presser, P. A. Ruud, V. K. Smith *et al.*, “The impact of" no opinion" response options on data quality: non-attitude reduction or an invitation to satisfice?” *Public Opinion Quarterly*, vol. 66, no. 3, pp. 371–403, 2002.

[56] F. J. Fowler, “How unclear terms affect survey data,” *Public Opinion Quarterly*, vol. 56, no. 2, pp. 218–231, 1992.

[57] D. L. Paulhus, *The role of Constructs in Psychological and Educational Measurement*. Psychology Press, 2002, ch. Socially desirable responding: The evolution of a construct, pp. 49–69.

[58] H. Van Herk, Y. H. Poortinga, and T. M. Verhallen, “Response styles in rating scales evidence of method bias in data from six EU countries,” *Journal of Cross-Cultural Psychology*, vol. 35, no. 3, pp. 346–360, 2004.

[59] H. Baumgartner and J.-B. E. Steenkamp, “Response styles in marketing research: a cross-national investigation,” *Journal of Marketing Research*, vol. 38, no. 2, pp. 143–156, 2001.

[60] B. Weijters, N. Schillewaert, and M. Geuens, “Assessing response styles across modes of data collection,” *Journal of the Academy of Marketing Science*, vol. 36, no. 3, pp. 409–422, 2008.

[61] B. Weijters, M. Geuens, and N. Schillewaert, “The stability of individual response styles.” *Psychological Methods*, vol. 15, no. 1, p. 96, 2010.

[62] P. B. Smith, “Acquiescent response bias as an aspect of cultural communication style,” *Journal of Cross-Cultural Psychology*, vol. 35, no. 1, pp. 50–61, 2004.

[63] J. G. Bachman and P. M. O’Malley, “Yea-saying, nay-saying, and going to extremes: black-white differences in response styles,” *Public Opinion Quarterly*, vol. 48, no. 2, pp. 491–509, 1984.

[64] J. A. Cote and M. R. Buckley, “Estimating trait, method, and error variance: generalizing across 70 construct validation studies,” *Journal of Marketing Research*, vol. 24, pp. 315–318, 1987.

[65] L. J. Williams, J. A. Cote, and M. R. Buckley, “Lack of method variance in self-reported affect and perceptions at work: reality or artifact?” *Journal of Applied Psychology*, vol. 74, no. 3, p. 462, 1989.

[66] J. A. Cote and M. R. Buckley, “Measurement error and theory testing in consumer research: an illustration of the importance of construct validation,” *Journal of Consumer Research*, vol. 14, pp. 579–582, 1988.

[67] J. D. Winkler, D. E. Kanouse, and J. E. Ware, “Controlling for acquiescence response set in scale development.” *Journal of Applied Psychology*, vol. 67, no. 5, p. 555, 1982.

[68] M. G. De Jong, J.-B. E. Steenkamp, J.-P. Fox, and H. Baumgartner, “Using item response theory to measure extreme response style in marketing research: a global investigation,” *Journal of Marketing Research*, vol. 45, no. 1, pp. 104–115, 2008.

[69] E. A. Greenleaf, “Measuring extreme response style,” *Public Opinion Quarterly*, vol. 56, no. 3, pp. 328–351, 1992.

[70] G. E. Lenski and J. C. Leggett, “Caste, class, and deference in the research interview,” *American Journal of Sociology*, vol. 65, pp. 463–467, 1960.

[71] B. Weijters, M. Geuens, and N. Schillewaert, “The proximity effect: the role of inter-item distance on reverse-item bias,” *International Journal of Research in Marketing*, vol. 26, no. 1, pp. 2–12, 2009.

[72] A. J. Arce-Ferrer, “An investigation into the factors influencing extreme-response style improving meaning of translated and culturally adapted rating scales,” *Educational and Psychological Measurement*, vol. 66, no. 3, pp. 374–392, 2006.

[73] C. Chen, S.-y. Lee, and H. W. Stevenson, “Response style and cross-cultural comparisons of rating scales among East Asian and North American students,” *Psychological Science*, vol. 6, no. 3, pp. 170–175, 1995.

[74] G. Marin, R. J. Gamba, and B. V. Marin, “Extreme response style and acquiescence among hispanics the role of acculturation and education,” *Journal of Cross-Cultural Psychology*, vol. 23, no. 4, pp. 498–509, 1992.

[75] G. Pell, “Use and misuse of Likert scales,” *Medical Education*, vol. 39, no. 9, pp. 970–970, 2005.

[76] R. Nickerson, “Null hypothesis significance testing: a review of an old and continuing controversy.” *Psychological Methods*, vol. 5, no. 2, p. 241, 2000.

[77] W. W. Tryon, “Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests.” *Psychological Methods*, vol. 6, no. 4, p. 371, 2001.

[78] H. Haller and S. Krauss, “Misinterpretations of significance: a problem students share with their teachers,” *Methods of Psychological Research Online*, vol. 7, no. 1, pp. 1–20, 2002.

[79] R. Hubbard, “Alphabet soup: blurring the distinctions between p’s and α’s in psychological research,” *Theory & Psychology*, vol. 14, no. 3, pp. 295–327, 2004.

[80] F. L. Schmidt, “Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers.” *Psychological Methods*, vol. 1, no. 2, p. 115, 1996.

[81] S. Goodman, “A dirty dozen: twelve p-value misconceptions,” in *Seminars in hematology*, vol. 45, no. 3, 2008, pp. 135–140.

[82] A. Stang, C. Poole, and O. Kuss, “The ongoing tyranny of statistical significance testing in biomedical research,” *European Journal of Epidemiology*, vol. 25, no. 4, pp. 225–230, 2010.

[83] R. L. Hagen, “In praise of the null hypothesis statistical test.” *American Psychologist*, vol. 52, pp. 15–24, 1997.

[84] J. Cohen, “The earth is round (p <. 05).” *American Pysychologist*, vol. 49, no. 12, p. 997, 1994.

[85] J. L. Rodgers, “The epistemology of mathematical and statistical modeling: a quiet methodological revolution.” *American Psychologist*, vol. 65, no. 1, pp. 1–12, 2010.

[86] T. Nix and J. Barnette, “The data analysis dilemma: ban or abandon. a review of null hypothesis significance testing,” *Research in the Schools*, vol. 5, no. 2, pp. 3–14, 1998.

[87] J. E. McLean and J. M. Ernest, “The role of statistical significance testing in educational research,” *Research in the Schools*, vol. 5, no. 2, pp. 15–22, 1998.

[88] G. Gigerenzer, “Mindless statistics,” *The Journal of Socio-Economics*, vol. 33, no. 5, pp. 587–606, 2004.

[89] E. Läärä, "Statistics: reasoning on uncertainty, and the insignificance of testing null," *Annales Zoologici Fennici*, vol. 46, no. 2, pp. 138–157, 2009.

[90] D. R. Anderson, K. P. Burnham, and W. L. Thompson, "Null hypothesis testing: problems, prevalence, and an alternative," *The journal of wildlife management*, vol. 64, pp. 912–923, 2000.

[91] R. Hubbard and R. M. Lindsay, "Why p values are not a useful measure of evidence in statistical significance testing," *Theory & Psychology*, vol. 18, no. 1, pp. 69–88, 2008.

[92] J. Cohen, "A power primer." *Psychological Bulletin*, vol. 112, no. 1, p. 155, 1992.

[93] M. R. Nester, "An applied statistician's creed," *Applied Statistics*, vol. 45, pp. 401–410, 1996.

[94] L. G. Daniel, "Statistical significance testing: a historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals," *Research in the Schools*, vol. 5, no. 2, pp. 23–32, 1998.

[95] P. R. Killeen, "An alternative to null-hypothesis significance tests," *Psychological Science*, vol. 16, no. 5, pp. 345–353, 2005.

[96] D. Bakan, "The test of significance in psychological research." *Psychological bulletin*, vol. 66, no. 6, p. 423, 1966.

[97] B. Thompson, "Statistical significance and effect size reporting: Portrait of a possible future," *Research in the Schools*, vol. 5, no. 2, pp. 33–38, 1998.

[98] D. Johnson, "The insignificance of statistical significance testing," *The Journal of Wildlife Management*, vol. 63, no. 3, pp. 763–772, 1999.

[99] R. Frick, "The appropriate use of null hypothesis testing." *Psychological Methods*, vol. 1, no. 4, p. 379, 1996.

[100] G. R. Loftus and M. E. Masson, "Using confidence intervals in within-subject designs," *Psychonomic Bulletin & Review*, vol. 1, no. 4, pp. 476–490, 1994.

[101] J. P. Ioannidis, "Why most published research findings are false," *Public Library of Science - Medicine*, vol. 2, no. 8, p. e124, 2005.

[102] J. D. Scargle, "Publication bias: the "file-drawer" problem in scientific inference," *Journal of Scientific Exploration*, vol. 14, no. 1, pp. 91–106, 2000.

[103] A. Stewart-Oaten, "Rules and judgments in statistics: three examples," *Ecology*, vol. 76, pp. 2001–2009, 1995.

[104] J. R. Levin, "What if there were no more bickering about statistical significance tests," *Research in the Schools*, vol. 5, no. 2, pp. 43–53, 1998.

[105] S. E. Maxwell, K. Kelley, and J. R. Rausch, "Sample size planning for statistical power and accuracy in parameter estimation," *Annual Review of Psychology*, vol. 59, pp. 537–563, 2008.

[106] R. E. McGrath, "Significance testing: is there something better?" *American Psychologist*, vol. 53, pp. 796–797, 1998.

[107] G. Cumming, F. Fidler, and D. L. Vaux, “Error bars in experimental biology,” *The Journal of Cell Biology*, vol. 177, no. 1, pp. 7–11, 2007.

[108] J. M. Hoenig and D. M. Heisey, “The abuse of power,” *The American Statistician*, vol. 55, no. 1, pp. 19–24, 2001.

[109] G. Cumming and S. Finch, “Inference by eye: Confidence intervals and how to read pictures of data.” *American Psychologist*, vol. 60, no. 2, p. 170, 2005.

[110] T. R. Knapp, “Comments on the statistical significance testing articles,” *Research in the Schools*, vol. 5, no. 2, pp. 39–41, 1998.

[111] P. Bacchetti, “Peer review of statistics in medical research: the other problem,” *British Medical Journal*, vol. 324, no. 7348, p. 1271, 2002.

[112] J. E. Bartlett, J. W. Kortlik, and C. C. Higgins, “Organizational research: determining appropriate sample size in survey research appropriate sample size in survey research,” *Information technology, learning, and performance journal*, vol. 19, no. 1, p. 43, 2001.

[113] *ITU-R BT.1788 : Methodology for the subjective assessment of video quality in multimedia applications*, International Telecommunications Union Std.

[114] G. Norman, “Likert scales, levels of measurement and the "laws" of statistics,” *Advances in Health Sciences Education*, vol. 15, no. 5, pp. 625–632, 2010.

[115] D. B. Judd, “The 1931 I.C.I standard observer and coordinate system for colorimetry,” *Journal of the Optical Society of America*, vol. 23, no. 10, pp. 359–373, 1933.

[116] K. Kelley, S. E. Maxwell, and J. R. Rausch, “Obtaining power or obtaining precision delineating methods of sample-size planning,” *Evaluation & the Health Professions*, vol. 26, no. 3, pp. 258–287, 2003.

[117] University of Düsseldorf. (2012) G*power 3. web page. University of Düsseldorf, institute of experimental psychology. Accessed 5.9.2012. [Online]. Available: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

[118] R. Luoto, “Kyselytutkimuksen suunnittelu,” *Duedecim*, vol. 125, pp. 1647–53, 2009.

[119] W. M. Kuzon Jr, M. G. Urbanchek, and S. McCabe, “The seven deadly sins of statistical analysis,” *Annals of Plastic Surgery*, vol. 37, no. 3, pp. 265–272, 1996.

[120] C. VanVoorhis and B. Morgan, “Understanding power and rules of thumb for determining sample sizes,” *Tutorials in Quantitative Methods for Psychology*, vol. 3, no. 2, pp. 43–50, 2007.

[121] L. Thomas, “Retrospective power analysis,” *Conservation Biology*, vol. 11, no. 1, pp. 276–280, 1997.

[122] S. N. Goodman and J. A. Berlin, “The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results,” *Annals of Internal Medicine*, vol. 121, no. 3, pp. 200–206, 1994.

[123] L. Z. Garamszegi, "Comparing effect sizes across variables: generalization without the need for Bonferroni correction," *Behavioral Ecology*, vol. 17, no. 4, pp. 682–687, 2006.

[124] M. D. Moran, "Arguments for rejecting the sequential Bonferroni in ecological studies," *Oikos*, vol. 100, pp. 403–405, 2003.

[125] S. Nakagawa, "A farewell to bonferroni: the problems of low statistical power and publication bias," *Behavioral Ecology*, vol. 15, no. 6, pp. 1044–1045, 2004.

[126] J. Cohen, "Things I have learned (so far)," *American Psychologist*, vol. 45, no. 12, p. 1304, 1990.

[127] D. H. Johnson, "Statistical sirens: the allure of nonparametrics," *Ecology*, vol. 76, pp. 1998–2000, 1995.

[128] R. Blair and J. Higgins, "Comparison of the power of the paired samples t -test to that of Wilcoxon's signed-ranks test under various population shapes." *Psychological Bulletin*, vol. 97, no. 1, p. 119, 1985.

[129] M. J. Nanna and S. S. Sawilowsky, "Analysis of Likert scale data in disability and medical rehabilitation research." *Psychological Methods*, vol. 3, no. 1, p. 55, 1998.

[130] T. G. Gregoire and B. Driver, "Analysis of ordinal data to detect population differences." *Psychological Bulletin*, vol. 101, no. 1, p. 159, 1987.

[131] E. Schmider, M. Ziegler, E. Danay, L. Beyer, and M. Bühner, "Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption," *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, vol. 6, no. 4, p. 147, 2010.

[132] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, pp. 677–680, 1946.

[133] P. F. Velleman and L. Wilkinson, "Nominal, ordinal, interval, and ratio typologies are misleading," *The American Statistician*, vol. 47, no. 1, pp. 65–72, 1993.

[134] J. Gaito, "Measurement scales and statistics: resurgence of an old misconception." *Psychological Bulletin*, vol. 87, pp. 564–567, 1980.

[135] H. Thomas, "IQ, interval scales, and normal distributions." *Psychological Bulletin*, vol. 91, no. 1, p. 198, 1982.

[136] J. T. Townsend and F. G. Ashby, "Measurement scales and statistics: the misconception misconceived." *Psychological Bulletin*, vol. 96, pp. 394–401, 1984.

[137] A. Zand Scholten and D. Borsboom, "A reanalysis of Lord's statistical treatment of football numbers," *Journal of Mathematical Psychology*, vol. 53, no. 2, pp. 69–75, 2009.

[138] D. J. Hand, "Statistics and the theory of measurement," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 159, pp. 445–492, 1996.

[139] D. Stengel, M. Bhandari, and B. Hanson, *Trauma-Statistics and Data Management: A Practical Guide for Orthopedic Surgeons*. Thieme, 2010.

[140] E. Svensson, “Guidelines to statistical evaluation of data from rating scales and questionnaires,” *Journal of Rehabilitation Medicine*, vol. 33, no. 1, pp. 47–48, 2001.

[141] B. D. Zumbo and D. W. Zimmerman, “Is the selection of statistical methods governed by level of measurement?” *Canadian Psychology/Psychologie canadienne*, vol. 34, no. 4, p. 390, 1993.

[142] W. G. Jacoby, “Levels of measurement and political research: an optimistic view,” *American Journal of Political Science*, vol. 43, pp. 271–301, 1999.

[143] T. R. Knapp, “Treating ordinal scales as interval scales: an attempt to resolve the controversy,” *Nursing Research*, vol. 39, no. 2, pp. 121–123, 1990.

[144] U. Jakobsson, “Statistical presentation and analysis of ordinal data in nursing research,” *Scandinavian Journal of Caring Sciences*, vol. 18, no. 4, pp. 437–440, 2004.

[145] J. Michell, “Measurement scales and statistics: a clash of paradigms.” *Psychological Bulletin*, vol. 100, no. 3, p. 398, 1986.

[146] W. W. Stine, “Meaningful inference: The role of measurement in statistics.” *Psychological Bulletin*, vol. 105, no. 1, pp. 147–155, 1989.

[147] E. Kahler, A. Rogausch, E. Brunner, and W. Himmel, “A parametric analysis of ordinal quality-of-life data can lead to erroneous results,” *Journal of Clinical Epidemiology*, vol. 61, no. 5, pp. 475–480, 2008.

[148] U. Jakobsson and A. Westergren, “Statistical methods for assessing agreement for ordinal data,” *Scandinavian Journal of Caring Sciences*, vol. 19, no. 4, pp. 427–431, 2005.

[149] U. Munzel and F. Langer, “A global view on parametric and nonparametric approaches to the analysis of ordered categorical data,” *Biometrical Journal*, vol. 46, no. 1, pp. 7–18, 2004.

[150] M. R. Harwell and G. G. Gatti, “Rescaling ordinal data to interval data in educational research,” *Review of Educational Research*, vol. 71, no. 1, pp. 105–131, Spring 2001.

[151] R. M. O’Brien, “The use of pearson’s r with ordinal data,” *American Sociological Review*, vol. 44, pp. 851–857, 1979.

[152] J. L. Rasmussen, “Analysis of Likert-scale data: a reinterpretation of Gregoire and Driver.” *Psychological Bulletin*, vol. 105, pp. 167–170, 1989.

[153] E. Abascal and V. D. d. Rada, “Analysis of 0 to 10-point response scales using factorial methods: a new perspective,” *International Journal of Social Research Methodology*, vol. 16, no. ahead-of-print, pp. 1–16, 2013.

[154] J. C. Hobart, S. J. Cano, J. P. Zajicek, and A. J. Thompson, “Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations,” *The Lancet Neurology*, vol. 6, no. 12, pp. 1094–1105, 2007.

[155] F. M. Lord, "On the statistical treatment of football numbers." *American Psychologist*, vol. 8, pp. 750–751, 1953.

[156] S. Boissard and M. Fontoynont, "Optimization of LED-based light blendings for object presentation," *Color Research & Application*, vol. 34, no. 4, pp. 310–320, 2009.

[157] N. Pousset, G. Obein, and A. Razet, "Visual experiment on LED lighting quality with color quality scale colored samples," in *Proceedings of CIE 2010 Lighting Quality and Energy Efficiency, 14-17 March 2010, Vienna, Austria. CIE: 722 -729.*, 2010.

[158] N. Sándor and J. Schanda, "Visual colour rendering based on colour difference evaluations," *Lighting Research & Technology*, vol. 38, no. 3, p. 225, 2006.

[159] K. Smet, W. Ryckaert, G. Deconinck, and P. Hanselaer, "A colour rendering metric based on memory colours (MCRI)," in *The CREATE 2010 Conference Proceedings*, 2010, pp. 354–356.

[160] F. Szabó, J. Schanda, P. Bodrogi, and E. Radkov, "A comparative study of new solid state light sources," in *Proceedings of 26th Session of the CIE*, vol. 1, 2007.

[161] F. Szabó, P. Bodrogi, and J. Schanda, "A colour harmony rendering index based on predictions of colour harmony impression," *Lighting Research & Technology*, vol. 41, no. 2, p. 165, 2009.

[162] F. Vienot, E. Mahler, J. Ezrati, C. Boust, A. Rambaud, and A. Bricoune, "Color appearance under LED illumination: the visual judgment of observers," *Journal of Light & Visual Environment*, vol. 32, no. 2, pp. 208–213, 2008.

[163] P. R. Boyce, J. A. Veitch, G. R. Newsham, C. Jones, J. Heerwagen, M. Myer, and C. Hunter, "Lighting quality and office work: two field simulation experiments," *Lighting Research & Technology*, vol. 38, no. 3, pp. 191–223, 2006.

[164] J. Bullough, K. S. Hickcox, T. Klein, and N. Narendran, "Effects of flicker characteristics from solid-state lighting on detection, acceptability and comfort," *Lighting Research & Technology*, vol. 43, no. 3, pp. 337–348, 2011.

[165] D. Carter and M. Al Marwaee, "User attitudes toward tubular daylight guidance systems," *Lighting Research & Technology*, vol. 41, no. 1, pp. 71–88, 2009.

[166] Y. De Kort and K. Smolders, "Effects of dynamic lighting on office workers: first results of a field study with monthly alternating settings," *Lighting Research & Technology*, vol. 42, no. 3, pp. 345–360, 2010.

[167] S. Fotios and C. Cheal, "The effect of a stimulus frequency bias in side-by-side brightness ranking tests," *Lighting Research & Technology*, vol. 40, no. 1, pp. 43–54, 2008.

[168] ——, "Brightness matching with visual fields of different types," *Lighting Research & Technology*, vol. 43, no. 1, pp. 73–85, 2011.

[169] L. Izsó, E. Láng, L. Laufer, S. Suplicz, and Á. Horváth, “Psychophysiological, performance and subjective correlates of different lighting conditions,” *Lighting Research & Technology*, vol. 41, pp. 349–360, 2009.

[170] E. Jaén, E. Colombo, and C. Kirschbaum, “A simple visual task to assess flicker effects on visual performance,” *Lighting Research & Technology*, vol. 43, pp. 457–471, 2011.

[171] H. Juslén, M. Wouters, and A. Tenner, “Preferred task-lighting levels in an industrial work area without daylight,” *Lighting Research & Technology*, vol. 37, no. 3, pp. 219–231, 2005.

[172] C. Knight, “Field surveys of the effect of lamp spectrum on the perception of safety and comfort at night,” *Lighting Research & Technology*, vol. 42, no. 3, pp. 313–329, 2010.

[173] K. Smet, W. Ryckaert, M. R. Pointer, G. Deconinck, and P. Hanselaer, “Optimization of colour quality of LED lighting with reference to memory colours,” *Lighting Research & Technology*, vol. 44, no. 1, pp. 7–15, 2012.

[174] R. Baniya, R. Dangol, P. Bhusal, A. Wilm, E. Baur, M. Puolakka, and L. Halonen, “User-acceptance studies for simplified light-emitting diode spectra,” *Lighting Research & Technology*, vol. 46, 2013.

[175] CIE. (2006) CIE statement on the recognition of journals in the field of light and lighting. CIE NEws 78. CIE.

[176] G. Newsham, M. Aries, S. Mancini, and G. Faye, “Individual control of electric lighting in a daylit space,” *Lighting Research &Technology*, vol. 40, no. 1, pp. 25–41, 2008.

[177] S. Jost-Boissard, M. Fontoynont, and J. Blanc-Gonnet, “Perceived lighting quality of LED sources for the presentation of fruit and vegetables,” *Journal of Modern Optics*, vol. 56, no. 13, pp. 1420–1432, 2009.

[178] F. Vienot, M. Durand, and E. Mahler, “The effect of LED lighting on performance, appearance and sensations: CIE light and lighting conference, budapest,” in *CIE Light and Lighting Conference with Special Emphasis on LEDs and Solid State Lighting, 27-29 May, 2009, Budapest, Hungary.*, vol. 2009, 2009, pp. 19–20.

Aalto-DD 15/2015

Investigating subjective aspects of lighting, such as preference and acceptance, involves evaluations and ratings by human subjects. The data from the subjects is gathered with questionnaire techniques. These methods are similar to those on other fields studying subjective impressions and evaluations, such as psychology, social sciences or marketing research. The results may be influenced by various factors, such as response styles of the subjects. There are also some controversy on how the data from questionnaires should be analysed.

9 789526 060651

ISBN 978-952-60-6065-1 (printed)
ISBN 978-952-60-6066-8 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Electrical Engineering and Automation
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

**DOCTORAL
DISSERTATIONS**