Better Images – Understanding and Measuring Subjective Image-Quality

Jenni Radun



Visual Cognition Research Group
Institute of Behavioural Sciences
University of Helsinki, Finland

Academic dissertation to be publicly discussed,
by due permission of the Faculty of Behavioural Sciences
at the University of Helsinki
in Auditorium A132 at the Institute of Behavioural Sciences, Siltavuorenpenger 1A,
on the 2nd of December, 2016, at 12 o'clock

University of Helsinki
Institute of Behavioural Sciences
Studies in Psychology 123: 2016

Supervisors

Docent Jukka Häkkinen, PhD, Institute of Behavioural Sciences, University of Helsinki, Finland

Professor Emeritus Göte Nyman, PhD Institute of Behavioural Sciences, University of Helsinki, Finland

Reviewers

Professor Patrick Le Callet, PhD, Polytech Nantes/Université de Nantes
IRCCYN/IVC, France

Associate Professor Marius Pedersen, PhD, Norwegian University of Science and Technology, Gjøvik, Norway

Opponent

Professor Ingrid Heynderickx, PhD, Eindhoven University of Technology, Netherlands

ISSN 0781-8254
ISSN-L 1798-842X
ISBN 978-951-51-2669-6 (pbk.)
ISBN 978-951-51-2670-2 (PDF)
http://www.ethesis.helsinki.fi
Helsinki University Print
Helsinki 2016

Contents

Abstract **6**

Tiivistelmä **8**

Acknowledgements **10**

List of original publications **12**

Abbreviations **13**

Glossary **14**

1 Introduction **15**

1.1 The challenge of high-image-quality estimation 16

1.2 What is image-quality? 17

1.3 Why is the estimation of image-quality difficult? 18

1.3.1 Images in subjective image-quality estimation 18

1.3.2 Artefactual image attributes 19

1.3.3 Preferential image attributes 20

1.3.4 Multiple attributes 20

1.4 The process of estimating image-quality 23

1.4.1 The general functioning of the visual system and attention 23

1.4.2 Material-related influences on estimations of image-quality 26

1.4.3 Eye movements in a quality-estimation task 27

1.4.4 Image-quality estimation as a preference task 29

1.5 Measuring subjective image-quality 30

1.5.1 Test-subject requirements 31

1.5.2 Test-material requirements 32

1.5.3 Test-condition requirements 34

1.5.4 Standard methods for subjective image-quality assessment 35

1.5.5 Qualitative methods 39
1.5.6 Behavioural and psychophysical registration 40
2 Research questions and hypotheses 41
3 Methods 44
3.1 Participants 44
3.2 Viewing conditions 44
3.3 Eye tracking 45
3.4 Qualitative analysis 45
3.5 Quantitative analysis 46
3.6 Eye-movement data analysis 47
4 Experiments and results 48
4.1 Study 1: Can naïve participants say on what they base their quality estimations? 48
4.1.1 Stimuli 49
4.1.2 Procedure 49
4.1.3 Results 50
4.2 Study 2: How non-trained estimators characterise the dimensions of image-quality? 55
4.2.1 Stimuli 55
4.2.2 Procedure 56
4.2.3 Results 57
4.3 Study 3: Do small changes in instructions change the way people seek information from images? 60
4.3.1 Stimuli 61
4.3.2 Procedure 62
4.3.3 Analyses 63
4.3.4 Results 64

4.4 Study 4: Are individual differences in viewing behavior related to different estimation rules in a quality-estimation task? 68
4.4.1 Experiment 1: Stimuli 68
4.4.2 Experiment 1: Procedure 69
4.4.3 Experiment 1: Analyses 70
4.4.4 Experiment 1: Results 70
4.4.5 Experiment 2: Introduction 74
4.4.6 Experiment 2: Stimuli 74
4.4.7 Experiment 2: Procedure 75
4.4.8 Experiment 2: Analyses 76
4.4.9 Experiment 2: Results 77
5 Discussion 80
5.1 The measurement of image-quality 80
5.1.1 Interpretation-based quality – the IBQ method 81
5.1.2 Eye-tracking in image-quality estimation 82
5.2 The process of quality estimation 84
5.2.1 Estimation rules 84
5.2.2 Context dependency 86
5.2.3 Subjectivity and individual differences 88
5.2.4 The process of visual high-image-quality estimation 90
5.3 Limitations 93
5.4 Recommendations for researchers conducting studies on subjective image-quality estimation 95
5.5 Conclusions 96
6 References 98

Abstract

The objective in this thesis was to examine the psychological process of image-quality estimation, specifically focusing on people who are naïve in this respect and on how they estimate high-quality images. Quality estimation in this context tends to be a preference task, and to be subjective. The aim in this thesis is to enhance understanding of viewing behaviour and estimation rules in the subjective assessment of image-quality. On a more general level, the intention is to shed light on estimation processes in preference tasks.

An Interpretation-Based Quality (IBQ) method was therefore developed to investigate the rules used by naïve participants in their quality estimations. It combines qualitative and quantitative approaches, and complements standard methods of image-quality measurement. The findings indicate that the content of the image influences perceptions of its quality: it influences how the interaction between the content and the changing image features is interpreted (Study 1). The IBQ method was also used to create three subjective quality dimensions: naturalness of colour, darkness and sharpness (Study 2). These dimensions were used to describe the performance of camera components. The IBQ also revealed individual differences in estimation rules: the participants differed as to whether they included interpretation of the changes perceived in an image in their estimations or whether they just commented on them (Study 4).

Viewing behaviour was measured to enable examination of the task properties as well as the individual differences. Viewing behaviour was compared in two tasks that are commonly used in studies on image-quality estimation: the estimation of difference and the estimation of difference in quality (Study 3). The results showed that viewing behaviour differed even in two magnitude-estimation tasks with identical material. When they were estimating quality the participants concentrated mainly on the semantically important areas of the image, whereas in the difference-estimation task they also examined wider areas. Further examination of quality-estimation task revealed individual differences in the viewing behaviour and in the importance these viewing behaviour groups attached to the interpretation of changes in their estimations (Study 4). It seems

that people engaged in a subjective preference-estimation task use different estimation rules, which is also reflected in their viewing behaviour.

The findings reported in this thesis indicate that: 1) people are able to describe the basis of their quality estimations even without training when they are allowed to use their own vocabulary; 2) the IBQ method has the potential to reveal the rules used in quality estimation; 3) changes in instructions influence the way people search for information from the images; and 4) there are individual differences in terms of rules and viewing behaviour in quality-estimation tasks.

Tiivistelmä

Tämän väitöskirjan tarkoituksena on tarkastella kuvanlaadun arviointia psykologisena prosessina, erityisesti miten kuvanlaadun arvioinnin suhteen naiivit koehenkilöt arvioivat korkealaatuisia kuvia. Laadun arviointi tällaisissa tapauksissa on usein preferenssi tehtävä, ja siten subjektiivinen. Tämän väitöskirjan tarkoituksena on lisätä tietoa subjektiivisen kuvanlaadun arviointitehtävän katselukäyttäytymisestä ja arviointisäännöistä. Yleisempänä päämääränä on ymmärtää preferenssitehtävien arviointiprosessia.

Tulkinnallisen laadun menetelmä (Interpretation-Based Quality method, IBQ) kehitettiin naiivien koehenkilöiden laatuarvioinneissaan käyttämien sääntöjen tarkasteluun. Menetelmässä yhdistetään laadullista ja määrällistä lähestymistapaa ja se täydentää perinteisiä kuvanlaadun mittausmenetelmiä. Tulokset osoittavat, että kuvan sisältö vaikuttaa sen laadun kokemiseen: sisällön ja kuvapiirteiden välinen yhteisvaikutus määrää miten kuvanlaatu tulkitaan (tutkimus 1). Tulkinnallisen laadun menetelmän avulla muodostettiin myös subjektiivisen kuvanlaadun kolme ulottuvuutta: luonnollisuus, tummuus ja tarkkuus (tutkimus 2). Näitä käytettiin kuvaamaan kameran komponenttien suoritusta. Tulkinnallisen laadun menetelmä paljasti myös yksilöiden välisiä eroja arviointisäännöissä: Koehenkilöt erosivat toisistaan siinä huomioivatko he arvioissaan vain kuvanlaatupiirteissä tapahtuneet muutokset vai myös miten nämä muutokset vaikuttivat kuvan tulkintaan (tutkimus 4).

Tehtävän ymmärrystä ja siinä ilmeneviä yksilöiden välisiä eroja selvennettiin katselukäyttäytymisen tarkastelun avulla. Katselukäyttäytymistä vertailtiin kahdessa yleisesti kuvanlaadun arvioinneissa käytetyssä tehtävässä: erojen ja laadun arvioinnissa (tutkimus 3). Tulokset osoittavat, että myös näissä kahdessa havainnon suuruuden arviointitehtävässä katselukäyttäytyminen oli erilaista, myös materiaalin ollessa identtistä. Laatua arvioidessaan koehenkilöt keskittyivät lähinnä semanttisesti tärkeisiin kuva-alueisiin, kun eroja arvioitaessa koehenkilöt tarkastelivat laajempia alueita. Laadunarviointitehtävän tarkastelu paljasti myös yksilöiden välisiä eroja sekä katselukäyttäytymisessä että säännöissä, joilla katselukäyttäytymisryhmät arvioivat kuvia (tutkimus 4).

Subjektiivisia preferenssiarvioita tehdessään ihmiset käyttävät erilaisia arviointisääntöjä, jotka näkyvät myös katselukäyttäytymisessä.

Tässä väitöskirjassa raportoidut tulokset osoittavat että 1) Ihmiset pystyvät perustelemaan laatuarvionsa myös ilman koulutusta, kun he saavat käyttää omaa sanastoaan; 2) Tulkinnallisen laadun menetelmä pystyy paljastamaan laatuarvioinneissa käytetyt säännöt; 3) Ohjeistuksen muutokset vaikuttavat siihen miten ihmiset etsivät tietoa kuvista; 4) Kuvanlaatua arvioitaessa yksilöiden välillä on eroja sekä arviointisäännöissä että katselukäyttäytymisessä.

Acknowledgements

First, I would like to thank my opponent and the two pre-examiners for their comments and for improving the quality of my work. They are the real experts in the field and understand the contribution of this doctoral thesis. However, they only see the final result, which in my opinion does not reveal the whole process.

The writing process has been long. First, I thank my mentor Göte Nyman, without whom I would not have started this process. I thank Göte for the inspiring environment he created for our group, which took us a long way. At the end of this process, my other mentor Jukka Häkkinen was more involved and it might be that without him I would not have finished the thesis. Jukka taught me a lot about scientific thinking. In addition to my mentors, I have had a lot of support from my colleagues, sharing ideas and anxieties. Tuomas Leisti and Toni Virtanen have been there for me throughout the whole process. Without them our research group would not have been the same. I am also grateful to Terhi Mustonen, Mikko Nuutinen and Oskari Salmi for their membership of our group.

Outside the University of Helsinki, I would like to thank the people who introduced me to the topic. They were working for Nokia at the time and we started our long-term collaboration in high spirits. I am particularly grateful to Tero Vuori, Mikko Vaahteranoksa and Jean-Luc Olives. In addition, Sebastian Arndt was also an inspiration for me during our international collaboration, which the Doctoral School of User-Centered Technology (UCIT) made possible. I would like to thank UCIT for financing my thesis, and also for creating a good environment in which to start my scientific career. Other organisations have also supported my research financially and I am grateful to the following: the Emil Aaltonen Foundation, The Ella and Georg Ehrnrooth Foundation, Kordelin Foundation and Nokia Foundation.

I wish to thank my parents for instilling persistence in me and teaching me that if you really want to do something you can if you put in the effort. Finally, I come to the man without whom I would not have started, continued or finished this process, my husband Igor. You have always challenged me and often attached more importance to this goal than I did. I also owe thanks to my children Milan,

Mira and now Meri who have shown me what life is all about –it is about caring for one another and showing that nothing is more important than being there when you are needed.

Helsinki, October 12th, 2016

Jenni Radun

List of original publications

Study 1 Radun, J., Leisti, T., Häkkinen, J., Ojanen, H., Olives, J. L., Vuori, T., & Nyman, G. (2008). Content and Quality: Interpretation-Based Estimation of Image-quality. *ACM Transactions on Applied Perception*, *4*(4), 1-15.

Study 2 Radun, J., Leisti, T., Virtanen, T., Vuori, T., Nyman, G., & Häkkinen, J. (2010). Evaluating the Multivariate Visual Quality Performance of Image-Processing Components. *ACM Transactions on Applied Perception*, *7*(3), 1–16.

Study 3 Radun, J., Leisti, T., Virtanen, T., Nyman, G., & Häkkinen, J. (2014). Why is quality estimation judgment fast? Comparison of gaze control strategies in quality and difference estimation tasks. *Journal of Electronic Imaging*, *23*(6), 061103.

Study 4 Radun. J., Nuutinen, M., Leisti, T., & Häkkinen, J. (2016). Individual differences in image quality estimations: Estimation rules and viewing strategies. *ACM Transactions on Applied Perception 13*(3), 1-22.

The articles are reprinted with the kind permission of the copyright holders.

Abbreviations

CA	Correspondence Analysis
EEG	Electroencephalography
FWHM	Full width at half maximum
FDM	Fixation density map
GEEs	Generalized estimating equations
GLMs	Generalized linear models
IBQ	Interpretation-based quality
ISO	the International Organization for Standardization
ISP	Image signal processor
ITU-R	The International Telecommunication Union's Radiocommunication sector
JND	Just-noticeable difference
JPG	Joint Photographic Experts Group's standard method for lossy compression of digital images
MOS	Mean Opinion Score
MTF	Modulation Transfer Function
PDF	Probability density function
rANOVA	Repeated measures analysis of variance
ROIs	Regions of Interest
sRGB	The standard RGB color space for monitors, printers and the internet.

Glossary

Abstract attributes	Attributes that are based on the interpretation of image features in specific image content.
Estimation rules	The set of attributes on which people base their estimations.
Feature-based attributes	Attributes that are based on the visibility of image features.
Image attributes	The subjective interpretation of image characteristics
Imaging devices	Devices that capture, process and represent images.
Image features	Characteristics in an image that can be objectively defined.
Memory colour	Colours that are recalled in association with familiar objects such as skin, grass and sky.
Objective measures	Measurements that rely only on physical properties, with no interpretation of meaning.
Perceptual attributes	Image characteristics that an observer senses.
Photospace distribution	The probability density function of the light levels and distances at which the photographs are taken.
Preference estimation	The subjective evaluation of superiority that arises from people's own experiences.
Quality experience	The entity that a person feels incorporates all the factors that influence quality, including the material as well as expectations and general preferences, for example.
Saliency models	Models based on image features predicting where a person would look.
Salient areas	Areas that are relevant from a strictly bottom-up perspective do not involve any interpretation of meaning.
Semantic Regions of the Interest (ROIs)	Areas that are relevant because of their significance to task.
Sensory evaluation	People's reports of object characteristics as they perceive them through their senses

1 Introduction

Try to remember how many (processed) images or videos you looked at yesterday. If the task is too difficult, try to estimate how many hours you spent in the process. Think of all the sources, your mobile phone, computer and tablet, magazines, newspapers, cameras, television, street advertisements, images in the supermarket and so on. We are constantly surrounded by visual information in the media and on our imaging devices. *Imaging devices* are devices that capture, process and represent images. Capturing and representing them is a process with various stages that potentially introduce errors into the image. We normally soon notice if the quality of the image is not good. Most people are able to distinguish between worse and better images even if the quality of the devices and hence of the processed images has improved in recent years to the extent that we are now used to images of fairly high quality. Nevertheless, many people faced with two versions of an image can still soon say which is the better one, or is better suited to a certain webpage on the Internet. How this process works is more difficult to explain.

One needs to understand the process of quality estimation to understand how cognition and vision work, as well as to improve the quality of imaging devices and the visual world surrounding us. One might think that modern digital technology and intelligent computational methods have already cracked the secret of visual quality preference, but this has not happened yet. Many image-quality algorithms have been developed in attempts to model quality perception among human participants by directly computing it from the image information, often based on knowledge about the human vision system and its functions (see Chandler, 2013 for a review). Such algorithms typically estimate the quality based on different *image features*, which is the term I will use here for the characteristics in an image that can be objectively defined, such as sharpness, colour and contrast. Image-quality algorithms are considered objective in that the calculations rely on physical features and patterns, typically without any interpretation of their meaning.

Why do we still need subjective image-quality estimation when objective estimations are also possible? There are at least three reasons: 1) subjective estimations are considered the basic truth against which objective metrics must be developed; 2) differences in image features are small when the quality level is high; and 3) image-quality estimation tends to be a preference task when the quality level is high. My focus in this thesis is on points 2) and 3).

1.1 The challenge of high-image-quality estimation

Given the improvements in the quality of imaging devices and images, objective measures cannot rely solely on the visibility of image features because the differences in the artefacts attributable to imaging devices are typically small. Hence, detection of the artefacts and distinguishing them from the images no longer suffice for the computation of image-quality. For the purpose of this dissertation I define images as high-quality when a participant can discriminate and identify everything, discriminability and identifiability having been stated as the requirements of quality (Janssen & Blommaert, 2000). When these requirements are fulfilled the quality estimation may be more of a preference task. This "beauty contest" is what an end-user is faced with when choosing an imaging device in terms of how "beautiful" one image is compared to others (Engeldrum, 2004a). My aim in this thesis is to enhance understanding of this beauty contest and of the related subjective processes in the context of consumer photographs.

The evaluations in these "beauty contests" are not based on technology variables or physical image parameters, but on *perceptual attributes*, in other words the characteristics of an image that a person actually senses (Engeldrum, 2004b). In this case, therefore, quality estimation is not directly related to the physical parameters of the image, but is rather *preference estimation* - the subjective evaluation of image superiority - and arises from people's own experiences. It is known that preferences are context-sensitive and are constructed at the time of the choice (Warren, Mcgraw, & Van Boven, 2011). Familiar preferences are generally well defined, but even in such cases situational factors may cause deviation from the most frequent choice (Bettman, Luce, & Payne, 1998). For example, someone who normally chooses ice cream for dessert might, on a cold day, prefer hot chocolate with marshmallows. The context-

dependency and subjectivity are the reasons why preferences are considered difficult to measure.

Therefore, the challenge of understanding subjective quality formation lies especially in the subjectivity and context-dependency of the quality-experience process. For example, the visibility and the meaning of different (physical) image features change depending on the content of the image, the context in which it is evaluated, and the reason why the person is looking at the image, and there are even personal preferences that are not well reflected in objective measures of image-quality. This is the challenge facing anyone attempting to understand the estimation of high-quality material, and it is what this dissertation is about.

1.2 What is image-quality?

The roots of image-quality estimation lie deep in the history of psychology, starting from the measurement of perception. Weber initiated the systematic measurement of sensations in the 19th century, and his measurements were further refined by Fechner, the acknowledged founder of psychophysics who developed systematic scales of perception (Gescheider, 1985). Interestingly, Fechner is also known as the founder of experimental aesthetics, as he started measuring people's preferences for artwork (Boring, 1957). Psychophysical methods are at the root of image-quality estimation nowadays, such as in the measurement standards (ISO 20462-1, 2005; ITU-R BT.500-13, 2012), which I examine in more detail later (Chapter 1.4).

Image-quality has been defined in several ways. It is described as "the intergraded perception of the overall excellence of an image", for example (Engeldrum, 2004b). This type of definition usually comes to mind when the talk is about quality: something is better than something else based on someone's evaluation of his or her own perception. Image-quality has also been defined as "an impression of its (image's) merit or excellence, as perceived by a participant neither associated with the act of photography, nor closely involved with the subject matter depicted" (Keelan, 2002, p.9). Here the stress is on the objectivity of the participant in relation to the image content. Further, the quality does not come from the content of the depicted image, it comes from the successful replication of some neutral object given that the memories related to a personal

subject might bias the evaluation process away from the target of quality. For example, a picture of your beloved but deceased dog does not have to be perfect to be valuable to you because it might represent all the good times you had together. Keelan (2002) also points out that quality is not associated with the act of photography. Hence, image-quality does not take into account how successful the composition is, or even the relevance of the photographed subject. A third definition stresses the usefulness of the image in fulfilling the quality requirements of discriminability and identifiability (Janssen & Blommaert, 2000). Hence, the appropriate use of an image depends on the ability to discriminate the information in it and to identify the items depicted. This definition leaves out the quality of high-quality images, however, the aim here being to determine how excellence is defined if the basic image-quality requirements are fulfilled.

1.3 Why is the estimation of image-quality difficult?

Why it is so difficult to determine which of two different artefacts influences image-quality more, or which imaging device is better, even though it is easy to judge subjectively which one of two images is better. The answer is in the interaction between visual processing and the material. This challenge is evident when attempts are made to construct algorithms that model human estimations of image-quality. Chandler (2013) lists the problems faced by developers of such algorithms in his review. These include the variety of possible distortions, the interaction between the distortion and the material, the multivariate changes in image-quality, geometrical changes, and changes due to image enhancement. Before addressing the challenges attributable to the interaction between visual processing and the material, it is necessary to know about the material.

1.3.1 Images in subjective image-quality estimation

The material used in subjective estimations of image-quality comprises natural images, in other words images of things or scenes from everyday life (Tolhurst, 2013). These are used because the visual processes concerned are sufficiently complex and representative. Artificial and frequently uniform test-target patches

that are often used in objective quality measurements are not normally used in subjective image-quality estimation because they lack the interaction between the image content and visual processing. Natural test images are not merely a collection of features, but also convey meanings and messages to the participants. For example, changes in the colours of a uniform patch and physically identical changes in a natural image have very different subjective consequences, which also affect the quality interpretations. People easily notice changes in skin colour, for instance, especially if the change makes the person look ill. Colours that are recalled in association with familiar objects are called "*memory colours*", and people are usually consistent in defining them, although they tend to be more saturated than real-world colours (Bartleson, 1960). However, given that memory colours are related to familiar objects, it is necessary to take into account the environmental properties, hence these memory colours may vary depending on the geographical location in which a person lives, for example. Typical objects for which people give consistent naturalness ratings include skin, grass and sky (Yendrikhovskij, Blommaert, & de Ridder, 1999). This is just one example of the interaction between meanings and image features.

1.3.2 Artefactual image attributes

Possible distortions in the quality of an image may cause changes in many of its features. Some of these distortions are related to artefacts coming from imaging devices, and some to environmental factors. Image contents determine how the changes in features are perceived and interpreted. For the purposes of this thesis I refer to the subjective interpretation of image characteristics as *image attributes*. Such attributes differ in their influence on quality, and have been classified as preferential and artefactual (Keelan, 2002). The latter come from image processing and are not always visible, but if they are detected they decrease the quality of the image. Examples of artefactual attributes include a lack of sharpness, noisiness, redeye and a variety of digital artefacts such as compression. Some are based on global (e.g. compression) and others on local (e.g. packet loss) distortions (Engelke, Kaprykowsky, Zepernick, & Ndjiki-Nya, 2011). It is suggested that local are stronger than global distortions as attention attracters (Engelke et al., 2011). Changes in artefactual attributes may also be geometric,

such as the optical distortions attributable to the camera lenses, image-enhancement algorithms or the sharpening algorithm for making the edges sharper, which at the same time boosts noise in the image (Chandler, 2013).

The interaction between the distortion and the material may result in the material masking the distortion such that it is not visible in all areas of the image (Chandler, 2013). Noise is not easily seen in busy images with many details, for example, but it is easily distinguished in uniform image areas. Blur, on the other hand, is not perceived in the uniform areas but it is in the busy areas. Moreover, the perceived degradations may be more disturbing on some surfaces than on others. Figure 1 shows two images (used in Study 3), both of which have similar levels of noise added to the whole area. Noise is differently visible in different areas, and also on different surfaces. It may be considered more disturbing in the sky than on a wall, for example, even though it is visible in both. Hence, different features of the image are differently visible in the different areas as well as in the different contents.

1.3.3 Preferential image attributes

Preferential image attributes include colour balance, contrast, colourfulness and memory colour reproduction. They are always visible in the images, but their optimal value depends on the taste of the viewer as well as the content. Figure 2 presents a pair of images (used in Study 4) in which the differences are clear, but relate mainly to colour balance and are therefore preferential.

1.3.4 Multiple attributes

Multivariate changes add an element of challenge to the methods of psychophysics, which traditionally use material that is strictly controlled. In an ideal situation only one variable would change, or if two did their interaction would be the target of the study. However, it is common in image-quality estimation for many changes to happen at the same time, especially when the focus is on changes attributable to different devices. However, only if the changes are small in magnitude is it possible to calculate the common influence of the attributes on quality estimations by summing the influence of each one separately (Keelan, 2002). This has also been found in estimations of liking related to

changes in quality features (Tinio, Leder, & Strasser, 2011). However, if the quality difference of one attribute is large in magnitude, even modest differences in other attributes have little influence on the perceived quality of the image (Keelan, 2002).



Figure 1. Two images (from Study 3) with the same level of noise added to show that interpretation of how disturbing the noise is in the image depends on the areas as well as the content. The noise is clearly visible in the sky and on the wall, but not in areas with many small details.



Figure 2. An example of two images (from Study 4) with equally clear details, but possibly different interpretations: these two images differ the most in terms of colour balance.

These challenges give some indication of the interaction between the visual process and the material. They also clarify why it is necessary to use the same type of material in the estimations as in the final product, in other words natural

images. However, it is not enough only to use natural images: several different contents must be included to deal with, the restriction of interactions between the distortion and the material, or geometric differences, for example (see Chapter 1.5.2 for more on the selection of test images). Further challenges arise from the nature of quality estimation as a process, which I consider next.

1.4 The process of estimating image-quality

As implied in definitions of image-quality, it is not enough to be able to distinguish the items depicted in an image, it is also necessary fully to interpret the information it conveys. The well-known phrase "A picture is worth a thousand words" is indicative of the wealth of information to be found in a single image. What happens in this process of estimating image-quality in the light of all these meanings? In the following I describe the general functioning of the visual system and attention, and then discuss what is known about the processes of image-quality estimation and preference formation.

1.4.1 The general functioning of the visual system and attention

Seeing requires the gathering of information via the eyes. Only a small area in the middle of the visual field is accurate (0.3-2 ° of the visual angle), and the further the target is from the area in the middle, the less accurately it is perceived (Land, 2006). Eye movements are used to sample the world around, and even though perception seems continuous and whole, visual perception is constructed mainly of stops and jumps to the next place, known as fixations and saccades. There are other types of eye movements (see e.g. Land, 2006), but in the context of looking at still images these are the most relevant.

Viewing strategies are commonly measured in terms of fixation duration and saccade amplitude, which are shorter in visual search than in scene perception, for example (180-275 ms and 3 degrees in visual search and 260-330 ms and 4-5 degrees in scene perception) (Rayner, 2009). The processing per fixation is therefore simple in the search task: whether the target is there or not. However, it is important not to jump over the target, and to screen the whole image. What matters in scene perception is to fixate many aspects of important areas rather

than all the areas. The duration of fixation has been associated with the difficulty of scene processing (Henderson, Nuthmann, & Luke, 2013): the longer the fixation the deeper the processing tends to be (Holmqvist et al., 2011). However, the length of fixation could also be related to how interesting the content is, as well as to impaired clarity. In other words, fixations may be long if there is a lot of information to be retrieved from one place or if the information is difficult to obtain. However, gaze duration on one place (including several fixations) could be a better measure than the duration of single fixations in the assessment of viewing strategies in different tasks (Castelhano, Mack, & Henderson, 2009). The amplitude of saccades is related to task demands, workload, the stimulus and current cognitive processes: the more demanding and heavy the task is, the shorter the saccade amplitude (Holmqvist et al., 2011).

Although the participant's attention is not always where the fixation is, it is typically directed at the fixated location or the next location to be fixated (Henderson, 2007). Therefore, the fixated place is considered a good enough approximation of attention allocation. Attention determines which information coming through the senses can access conscious processing and working memory (Baddeley, 2003). Working memory maintains and stores information in the short term and underlies human thought processes, and is limited in nature (Baddeley, 2003).

Attention comprises bottom-up and top-down processes. Bottom-up attention refers to salience filters in the central nervous system that are selective for properties of stimuli that are likely to be important (Knudsen, 2007). These properties are easily distinguished, and include movement and differing colours and orientations. Objects with such properties pop out of the scene without any mental effort (see Treisman & Gelade, 1980). As Le Callet and Niebur (2013) suggest, I refer to areas that are relevant in a strictly bottom-up sense as "*salient*". Top-down mechanisms stem from the aims behind actions and regulate the signal strengths of different information channels that compete for access to the working memory (Knudsen, 2007). Such mechanisms direct the eye movements towards targets and improve the signal-to-noise ratio in all domains of information processing: sensory, motor, internal state and memory (Knudsen, 2007). They also direct the gaze to areas that are relevant to a certain action or

task and further make the detection of important features more sensitive than of the non-task-relevant features. The areas of attentive focus are relevant because of their meaning to the task, and the process relies on both bottom-up and top-down information. Le Callet and Niebur (2013) call these "important areas", but in this theses I refer to them as *semantic regions of interest (ROIs)* so as to emphasise the interpretation of bottom-up features that essentially distinguish between these salient and important areas. The meaning of information coming through the senses is thus constantly being processed. However, knowing about attention and eye movements does not in itself suffice to explain the process of quality estimation. It is also necessary to understand the cognitive processes that enable us to act in our environment and to interpret the things we perceive.

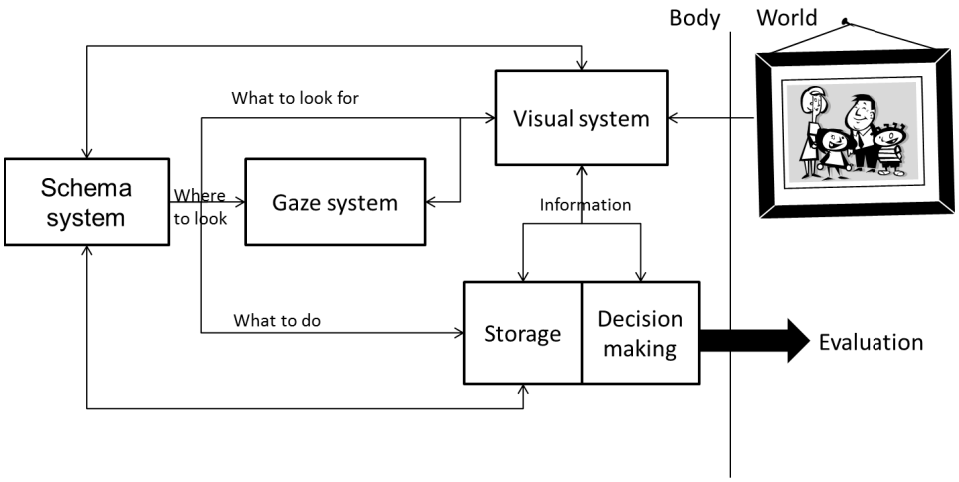


Figure 3. The flow of visual-quality estimation, modified from (Land, 2009)

Distinct components of the gaze-action system have been identified: schema control, the gaze system, the visual system and the motor system (Land, 2009) (Figure 3). Through these components individuals gather information from the outside world that they use to act in it. The gaze system serves to locate information thereby answering the question "where", whereas the visual system responds to the "what" question and supplies information on which to base action (Land, 2009). There are also different neural routes in the perceptual system. Land (2009) defines the schema system as determining where to look, what to

look for and what to do. Its role is twofold: setting the goal of the current behaviour and determining the sequence of actions needed to achieve it. Understanding a task requires an understanding of its schema: how the task should be done, what the important features are and how the decisions should be formed. I will now describe in more detail what happens in the interaction between a participant and images in a quality-estimation task.

1.4.2 Material-related influences on estimations of image-quality

Natural images are used as material in tasks related to image-quality estimation, and are also frequently used in research on attention allocation in other visual tasks. In the following, therefore, I consider the influence of bottom-up information on attention in general. Attention is differently allocated in simple and complex images: in the case of simple images with only a few attention catchers the participants watch the same places, whereas the fixations are widely distributed if complexity is high, as in noise images (Judd, Durand, & Torralba, 2011). Models have been developed to predict the salient places at which a person would be looking. Such models are based on image features, in other words on bottom-up information (Itti & Koch, 2000; Walther & Koch, 2006), and I refer to them as *saliency models*. They exploit knowledge about the functioning of the human visual system to predict attention allocation. For example, estimations of attention allocation in Itti and Koch's (2000) model are based on image features such as colours, orientation and intensity. However, these are not the only factors influencing where humans look. It was concluded from eye-movement data gathered among humans watching a large set of images (1,003) that people first look at text, other people and faces, and if none of these are present the attention is directed to the centre of the image and to the salient areas (Judd, Ehinger, Durand, & Torralba, 2009). Faces and text always draw people's attention, and these were incorporated into the newer version of Itti and Koch's model (Cerf, Frady, & Koch, 2009). The global context of an image also influences where people direct their gaze, and they look at different points in if there is a clear horizon than if there is an object in the middle, for example (Oliva & Torralba, 2007). This global context has been integrated into saliency models to stress the importance of salient areas depending on the global context: in a street view the

model concentrates on the salient areas below the horizon, for instance (Torralba, 2003).

However, recent studies have shown that saliency models work only in limited conditions, In such cases it is suggested that they work because objects are usually fixated on and usually they are salient (Einhäuser, Spain, & Perona, 2008). Therefore it is not the colours, orientations and intensity as such that direct the attention, but the need to recognise the objects. It has also been posited that it is cognitive relevance rather than low-level saliency that directs the attention (Henderson, Malcolm, & Schandl, 2009). In other words, fixated areas are selected based on the need of the cognitive system to understand the meaning of a scene in interaction with the goals of the current task. For example, if we look for a mug in the kitchen we do not start at the oven or the stove, even though they could be the most salient areas, we probably start with the shelves and countertops. The task requirements have been shown to reverse the effects of low-level saliency (Einhäuser, Rutishauser, & Koch, 2008), which is generally less relevant to attention allocation than the top-down influences, but of course both influence the allocation of attention.

1.4.3 Eye movements in a quality-estimation task

According to Land (2009), the schema system determines where we look, what we look for and what we do. Its role is twofold: to set the goal of the current behaviour and to determine the sequence of the actions needed to achieve it. It has been noted that task requirements influence eye-movement patterns (e.g. Castelhano et al., 2009; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011; Yarbus, 1967). For example, people engaged in active tasks such as visual search or reading use similar viewing strategies, which change when doing passive tasks such as watching in a dark room, or viewing a natural scene or simple patterns (Andrews & Coppola, 1999).

Eye-movement tracking is often used to estimate the allocation of attention in studies focusing on image-quality to improve the performance of objective quality metrics (Engelke et al., 2013; Larson, Vu, & Chandler, 2008; Liu & Heynderickx, 2011). Tasks that are frequently used to reveal the areas attended to include quality-estimation and free-viewing tasks, both of which appear to enhance the

performance of such metrics (Liu & Heynderickx, 2011). The quality-estimation task is used in experiments because it reflects what people normally do when estimating quality. On the other hand, the free-viewing task is thought to better capture the viewing behaviour of a normal end-user who would be looking at the final products, i.e. images. The maps of spatial-fixation density obtained from the free-viewing task have been found to improve objective metrics more than maps from the quality-estimation task (Larson et al., 2008).

The differences in attention allocation between quality-estimation and free-viewing tasks have been examined to some extent. In general, in the latter the fixations concentrate more on the most prominent regions of interest, whereas in the former the attention also wanders to other regions in search of cues to determine the level of image-quality (Alers, Redi, Liu, & Heynderickx, 2015). When the two tasks were compared, the globally distributed degradations (such as blurring and white noise) did not change the fixation allocation, but they did change it if the degradations were local (such as packet loss distortion, JPEG or JPEG2000) (Vu, Larson, & Chandler, 2008). In such cases the participants estimating quality tended to fixate more on the regions, where the degradations were visible, than those who were freely viewing images. It is not only the type of degradation, but also the contents that influence the gain achieved from adding the spatial-fixation distribution into objective image-quality metrics: the biggest improvements are in the contents in which the participants consistently fixated on the same image areas (Liu, Engelke, Le Callet, & Heynderickx, 2013). These kinds of contents have few clear, salient areas such as faces or text. There was less improvement in metrics in the contents with no clear attention catchers.

It is necessary to understand the schema of a task to understand the cognitive requirements. What are the requirements of an image-quality-estimation task? Earlier I defined the goal of this thesis: to examine image-quality estimation in the context of high-quality material, which is often a preference task. It is therefore necessary first to understand the special characteristics of a preference task. It has been noted that eye movements in a preference task differ from those in a free-viewing task, with shorter fixation durations at least at the beginning of the viewing, and longer saccade amplitudes (Mills et al., 2011). Preference tasks have not been extensively studied because of the inherent subjectivity.

Determining the requirements related to the preference task of quality estimation necessitates the expansion of investigations into gaze control and image-quality to the research realm of cognition and decision-making.

1.4.4 Image-quality estimation as a preference task

The special characteristic of a preference task is its subjectivity. We all have our own opinions. Preferences are sensitive to the context and are constructed at the time the choice is made (Warren et al., 2011). These aspects emphasise the psychological processes that are going on in the making of decisions or estimations, including consideration, weighting and valuation, and the integration of the relevant inputs (Warren et al., 2011). All this requires an understanding of relevant personal values (Payne, Bettman, & Schkade, 1999) as well as situational factors (Bettman et al., 1998).

The more that is known about the set of values built up in certain situations, the easier it is to comprehend subjectivity (Payne et al., 1999). People have different values, and another approach to subjectivity would be to examine the reasons behind individual differences. Individual differences in performance have been linked to computational limitations and differing construals of the task among the subjects (Stanovich & West, 2000). "Computational limitations" refer to differences in cognitive capacity that include, differences in working memory, for example (Bleckley, Durso, Crutchfield, Engle, & Khanna, 2003). "Different construals" of a task mean that people might understand it differently. Understanding a task in a certain way may lead to the use of specific, related deduction rules (Kruglanski & Gigerenzer, 2011). The rules on which people base their decisions in the case of visual-quality estimation reflect the set of reasons they consider important for that task. One person might estimate high-quality images according to the colours, whereas others may pay attention to sharpness. These rules are then reflected in their estimations as well as in the way they search for information.

It is common in studies investigating different aspects of image-quality to ask participants to assess images according to certain quality attributes such as sharpness, graininess, lightness and colour saturation (Virtanen, Nuutinen, Vaahteranoksa, Oittinen, & Häkkinen, 2015). This gives an indication of how

much the attributes disturb the quality. Are these the attributes they would use for all the image contents if the instruction did not direct their attention to them? People may well use different rules for their estimations, and these different rules may cause the large variations seen in preference tasks. On the other hand, if we knew the rules and could classify people into subgroups accordingly, for example, the variance would be reduced and the quality estimations related to certain material would be better understood. However, existing standards and recommendations concerning methods for estimating visual quality do not support this kind of examination. I explain the current standards and recommendations related to the subjective estimation of image-quality in the next section.

1.5 Measuring subjective image-quality

The standard methods of image-quality assessment come from the long tradition of psychophysics. The measurement of sensation dates back to 1834 when E. H. Weber noted that the differentiation of two relatively heavy weights required that they differ more than two relatively light weights (Boring, 1957). G.T. Fechner further refined Weber's work in calculating a scale of sensation magnitudes (Gescheider, 1985). The scale was based on the term *"just noticeable difference" (JND),* which Fechner used as a unit of sensation on a psychological scale that started at the absolute perceptual threshold. Fechner is considered the founder of psychophysics on account of this systematic measurement of sensation (Boring, 1957). Even nowadays JND is a commonly used measure of detectability that leads to the 75:25 proportion of responses in a task comparing two univariate stimuli, which are assessed in terms of a single attribute (e.g. ISO 20462-1, 2005). Fechner posited that sensation magnitude increases with the logarithm of stimulus intensity, but for this calculation it is necessary to know both the stimulus and the assessment measurements. This is not always possible.

In 1927, L.L. Thurstone developed methods for measuring sensory experience when the physical stimuli cannot be specified, the first psychologist to do so (Gescheider, 1985). He proposed that it was possible to calculate the psychological scale values for two stimuli from the proportion of times one was judged greater than the other with respect to a predefined attribute. Accordingly,

indirect measures of the ability to differentiate something were used to estimate sensation magnitude. The next step was taken in the 1950s when S.S. Stevens started asking people to directly assign a number to an observed stimulus that corresponded to the magnitude of the experienced sensation (Gescheider, 1985). This method of magnitude estimation replaced Fechner's logarithmic law. According to Stevens' power law, the estimated magnitude of sensory dimension increases in proportion to the stimulus intensity raised to a power, where the power exponent depends on the sensory modality and the stimulus conditions (Gescheider, 1985).

All these concepts are still applied today in the estimation of subjective image-quality. Paired comparison and magnitude estimation are commonly used (ISO 20462-1, 2005; ITU-R BT.500-13, 2012), and the JNDs have a key role in the new standard proposed for the subjective measurement of image-quality (ISO 20462-3, 2012). In the following sub-sections I describe common measurement techniques used for the subjective estimation of visual quality, and evaluate them from the perspective of estimation involving high-quality material. In the main I will go through the International Organization for Standardization's (ISO) recommendations that define the psychophysical experimental methods for estimating image-quality, as well as the recommendations of the International Telecommunication Union's Radiocommunication sector (ITU-R) with regard to methodology for the subjective assessment of the quality of television pictures, for example.

1.5.1 Test-subject requirements

According to the recommendations, participants must have normal vision, tested for visual acuity and colour vision (ISO 20462-1, 2005; ITU-R BT.500-13, 2012). They should be free from personal involvement in the design of the experiment as well as the subject matter depicted by the test stimuli (ISO 20462-1, 2005). Their expertise in image artefacts should be decided according to the objectives of the experiment: they may be experts or naïve (ITU-R BT.500-13, 2012). Expert observers should be used in critical studies, for example, whereas naïve observers are recommended in assessments of the quality of a final product. There should be at least 10 and preferably 20 subjects contributing to the analysis, and the

proportion of excluded subjects should not exceed 15 per cent (ISO 20462-1, 2005). If the number of participants is less than 15 the study is explorative, and should be referred to as informal (ITU-R BT.500-13, 2012). The way the participants are recruited as well as their level of expertise should be described.

1.5.2 Test-material requirements

The standards give several guidelines for the selection of test material depending on the purpose of the study, but they all require natural images (e.g. ISO 12640-1, 1997; ISO 20462-1, 2005; ITU-R BT.500-13, 2012). In the case of studies on image-quality the recommended minimum number of test images is three, but preferably six or more (ISO 20462-1, 2005). Content recommendations for the estimation of television pictures depend on the purpose of the study (ITU-R BT.500-13, 2012). In the case of overall performance estimation, for example, the images should be general and critical, but not unduly so, whereas the material should be critical in capacity and performance testing. The selected images should therefore be sensitive to problematic image artefacts. When the aim is to identify various imaging or image-transmission problems, the material should either be attribute-specific or wide-ranging and very rich, depending on the context. Another approach, introduced recently, is to use images selected according to eye-movement distributions (Farnand, 2013): it is recommended to use images with single points of focus for image-comparison purposes because the fixated places remain consistent among the participants. A further alternative is to use scenes with a uniform content. The rationale behind this approach is to prevent the effect of local feature changes such as hue and saturation shifts from altering the way the participant's attention is allocated, as can easily happen when looking at a busy picture.

There have also been attempts to define proper sets of test images for subjective image-quality estimation. The ISO has published several recommended image sets designed for this purpose, such as ISO 12640-1 (1997) and ISO 12640-2 (2004), and updates are available. These image sets are intended to measure the effects of different artefacts, showing for example skin tones and fine details as well as complicated geometric shapes (ISO 12640-1, 1997). However, when made sensitive to different image-quality artefacts these

sets are criticised for not representing the contents of everyday photography, which is important when testing imaging devices, and especially cameras.

One way of assessing what kind of images are commonly taken with cameras is to position them in a photographic space, or *photospace distribution*, which is the probability density function (PDF) of the light levels and distances at which the photographs are taken (Keelan, 2002; Segur, 2000). These two factors are outside the control of system designers, but influence the performance of imaging systems, especially cameras.

The photospace distribution collected from images taken with a compact point-and-shoot 35-mm-format camera shows two clear peaks: one with a moderate-to-long distance under bright light, corresponding to outdoor images during daylight, and another with a short-to-moderate distance in low light levels, primarily corresponding to indoor flash images (Keelan, 2002). The International Imaging Industry Association's (I3A) Camera Phone Image Quality (CPIQ) Initiative Group applied the photospace distribution obtained from the images taken with camera phones when they started to define guidelines for an image set to be used for testing the image-quality of camera phones (I3A, 2007). This distribution was weighted more towards the low lighting condition and short camera-subject distances than the distribution from compact point-and-shoot 35-mm-format cameras. Using the camera-phone photospace distribution as an estimate of camera-phone usage, the developers defined six clusters that encompass 70 per cent of images (I3A, 2007), which they recommended as guidelines for testing consumer experiences of camera-phone performance (Table 1).

In sum, there seems to be a consensus that natural images with several contents should be used as test material, but apart from that the recommendations vary or depend on the purpose of the study. The contents should be selected either to be sensitive to the artefact(s) under examination or to represent the types of images commonly produced with a certain device. Furthermore, natural images as such are complex stimuli, and it should be borne in mind that memory colours (e.g. of skin, sky and grass) matter in assessing the naturalness of colours, that different characteristics of the image influence the visibility of its artefacts, and that attention is differently distributed depending

on the content. Given the vast collection of recommendations and views on the selection of test images, careful reporting of the test material in each experiment is crucial.

Table 1. The test-image content clusters defined by i3a to represent ~ 70% of images taken with camera phones (reproduced from I3A (2007)).

Cluster	Subject Illuminance (Lux)	Subject-Camera Distance (m)	Typical Scene Description
1	< 50 Lux	~ 1 m	Close-up in dim-dark lighting conditions (indoor/outdoor)
2	50 -100 Lux	~ 1 m	Close-up in typical indoor lighting conditions (indoor/outdoor)
3	< 50 Lux	> 4 m	Small group in dim-dark lighting conditions (indoor/outdoor)
4	50 -100 Lux	> 4 m	Small group in typical indoor lighting conditions (indoor/outdoor)
5	> 3400 Lux	0.5 - 2 m	Small group in cloudy bright to sunny lighting conditions (outdoor)
6	> 3400 Lux	> 7 m	Scenic landscapes/ large groups in cloudy bright to sunny lighting conditions (outdoor)

1.5.3 Test-condition requirements

The viewing-condition requirements depend on the purpose of the research. The lighting should be higher (from 1500 lx to 2500 lx) for the critical than for the practical evaluation of print images, conforming more closely with common lighting levels at home or in the office (from 375 lx to 625 lx) (ISO 20462-1, 2005). The ITU defines different viewing conditions for testing television pictures in laboratory and home environments, including lighting and viewing conditions as well as the display settings (ITU-R BT.500-13, 2012). The colour settings used for coding images also influence the recommendations about monitor calibration as well as the viewing environment. One commonly used standard colour space is sRGB, which also includes recommendations covering colour calibration on the monitors as well as the viewing environment (defined, for example, in IEC 61996-2-1, 1999).

The duration of an experiment must be reasonable to prevent participant fatigue. One recommendation is that experiments, including giving instructions, should not exceed 45 minutes and must not exceed 60 minutes: if the experiment is longer the subjects should be given the opportunity to finish the test later (ISO 20462-1, 2005). Each test situation and method has its standards, which change according to the purpose of the study. The above are just a few examples of what should be considered and reported. Next I will introduce some common methods used in the estimation of visual quality.

1.5.4 Standard methods for subjective image-quality assessment

1.5.4.1 Paired comparison and the like

Paired comparison has been used as a method since the early days of psychophysics, following Weber's observation that the noticeable difference in weights depended on whether the weights were relatively light or heavy (Gescheider, 1985). The subject selects from two simultaneously presented images the one that fulfils a predetermined requirement, such as better image-quality or less of some image artefact (ISO 20462-1, 2005). Variations of the method include assessing the pairs on a comparison scale with either separate categories (such as *much worse, worse, slightly worse, the same, slightly better, better, much better*), or on a non-categorical scale defining only the ends and estimating the distance from them on a graphical scale or with numbers (ITU-R BT.500-13, 2012). A variation of this is to show two images on a display one after the other: the first one is shown, then the second one and then the first one again, after which the subject evaluates the difference between the image pair (ITU-R BT.500-13, 2012).

Because paired comparison is sensitive to small differences it can be used to determine JNDs. This is possible for some image attributes, such as sharpness and noise, or for general quality (ISO 20462-1, 2005). However, an attribute JND is not straightforward in that the contents of the image also influence how easily different image features are distinguished. The contents also influence quality JNDs: the JND distribution of responses is used to estimate the importance of quality variation, but this time in stimuli pairs that have multivariate changes and in terms of overall image-quality (ISO 20462-1, 2005).

Paired comparison is accurate when there are small differences, and is therefore good for assessing high-quality images. However, it is difficult to determine when assessing overall quality with multivariate differences which image attribute that is systematically changing is used as a criterion. This might also change from one participant to another. One weakness of paired comparison is the need for long and tiresome experiments, because all the images from a set should be assessed against all the others. Furthermore, paired comparison does not allow the reliable estimation of stimulus differences of more than 1.5 JNDs, because the response saturates (ISO 20462-1, 2005).

1.5.4.2 Rank ordering, categorical sorting and the like

Rank ordering means putting a set of images in order according to some rule, such as quality (ISO 20462-1, 2005). Categorical sorting, in turn, involves classifying the stimuli into one or several ordered categories, at least some of which are identified by adjectives or phrases that describe different levels of the attributes or image-quality (ISO 20462-1, 2005). A fair number of images may be used and these tasks are easy to understand. However, if the differences are small, as they often are at high levels of quality, the task may become difficult and rank ordering may not be sensitive enough as a method. In addition, the ratings are related to the selection of images in the set, and comparison between different sets may be somewhat difficult. One way round this is to use only a few same stimuli in both tests among other stimuli, which would then make comparison between the sets feasible. Another point is that the adjectival categories, even if ordered, cannot give the distances between the images because the distances between adjectival categories are not equal (ISO 20462-1, 2005). Rank ordering may also be difficult if the images to be assessed are large in size or presented on a display.

1.5.4.3 Magnitude estimation and the like

Magnitude estimation requires the participant to assign a numerical value to the stimulus that proportionally describes a predetermined attribute (Gescheider, 1985; ISO 20462-1, 2005). A reference stimulus or stimuli are usually presented to anchor the rating scales (ISO 20462-1, 2005). The scales may be numerical,

such an 11-grade categorical scale, or non-categorical, such as a graphical or numerical scale (ITU-R BT.500-13, 2012). In the case of graphical scales the subject assigns each image or image sequence a point on a line drawn between two semantic labels, and the distance from the end of the scale is used as a value. However, not even the steps in a graphical scale are equal if they are associated with different quality terms (Teunissen, 1996). Numerical scaling, in turn, requires the subject to assign a number that reflects the judged (subjective) level on a specific dimension. The range of numbers may be restricted, or if not then the task is to judge the level relative to that of the reference image.

Magnitude estimation is commonly used in studies on image-quality because it gives a single value describing the subjects' opinions, often termed a Mean Opinion Score (MOS). The term MOS is used in research on the quality of telephone transmission, for example, in which it is defined as "...the mean of values on a predefined scale that subjects assign to their opinion..." (ITU-T P.800.1, 2006). It has also been adapted for use in image-quality estimation. The MOS may be the subjects' estimation of the general quality or of the importance of a certain attribute in influencing quality. The number of stimuli may be considerably larger than with paired comparison, but magnitude estimation is not as accurate (ISO 20462-2, 2005), and it may also be somewhat difficult for an untrained subject.

The selection of references modifies the scale and the order of material presented influences the estimations, hence the stimuli have to be randomised and there must be enough subjects. The problem with reference selection is especially pronounced when the performance of imaging devices is being tested: then the variations are multivariate and each image is different, therefore no image is absolutely the best or the worst. One way round the problem is to introduce a dynamic reference: all the other images in the test set serve as reference images and are shown before each image estimation (Nuutinen et al., 2016). One limitation of this method is the need to restrict the number of test images, otherwise the experiment becomes too long.

1.5.4.4 Triplet comparison

The recommendation ISO 20462-2 (2005) introduced triplet comparison as a method that involves the simultaneous scaling of three test stimuli with respect to an image-quality attribute or overall quality. The aim is to achieve the same levels of accuracy and consistency as the paired-comparison method, but with less stress for the subjects given that the time required for triplet comparison is about one third of that needed for paired comparison (ISO 20462-2, 2005). If the test-image set is too large it is possible to combine a categorical step with triplet comparison, in which images of a similar quality level (e.g. favourable, acceptable or unacceptable) are classified in a common group. The comparisons are then made only within these groups, thereby reducing the required number. Triplet comparison is seen as a compromise between paired comparison and magnitude estimation – it is almost as accurate as the former, and almost as fast as the latter (ISO 20462-1, 2005).

1.5.4.5 The Quality Ruler method

A quality ruler is a reference-stimulus scale constructed from stimuli depicting the same scene with univariate manipulations that are arranged in the order of JNDs (ISO 20462-3, 2012). The quality ruler can be presented in either a hard-copy or a soft-copy format. The test stimuli are compared to this ruler and the image that most closely matches the test image provides the rating. Quality rulers can be made for attributes that are artefactual in nature, sharpness manipulation through the modification of the modulation transfer function (MTF) being common. They can also be used to estimate the differences in other types of attributes: a sharpness ruler can be used to estimate differences in colour tone, for example. The Quality Ruler method is suitable for measuring differences exceeding one JND, and gives an evaluation that is anchored against physical standards. (ISO 20462-3, 2012)

What is somewhat problematic is that the ruler must be defined solely in terms of artefactual attributes. According to the recommendation (ISO 20462-3, 2012), such rulers can also be used to measure other attributes against it. This would be the case if the amount of colour change were estimated against a sharpness ruler, the question then being where the degradation in quality is equal with the two

attributes. What really is measured in such cases is cumbersome. Furthermore, the requirement for different contents inhibits the flexible use of the ruler.

1.5.5 Qualitative methods

Sensory evaluation reflects a slightly different approach to quality estimation, in that humans are seen as "measurement instruments" reporting the characteristics of a product or material as they perceive them through their senses (Zeng, Ruan, & Koehl, 2008). The method was first developed to study the reactions of consumers to food products, but has been since used in many areas of quality inspection, product design and marketing (Zeng et al., 2008). It incorporates the use of classic psychophysical methods, as well as descriptive analysis to characterise the stimulus under examination on predefined scales or even in free descriptions of its properties (Civille & Oftedal, 2012; Meilgaard, Civille, & Carr, 1999). The properties are evaluated in terms of their quality, intensity or change over time (Meilgaard et al., 1999). The assessors tend to be trained or expert panellists with common training in how to characterise and assess certain stimuli, but naïve assessors are also used sometimes (Meilgaard et al., 1999). Another approach in which the assessment may be quantitative or qualitative is to measure consumer understanding, which requires naïve assessors (Civille & Oftedal, 2012). The focus in quantitative assessment is on the perceived intensity of certain characteristics, whereas qualitative assessment involves mapping the language of consumers and their emotions related to products and usage behaviours.

Panels defining concepts are sometimes also used for assessing image-quality (Bech et al., 1996), but consumer understanding among naïve participants tends to be restricted to assessments of general quality, preferences or estimations rated on different predefined scales. Free descriptions are rarely used to enhance consumer understanding in image-quality estimation, although they could provide valuable information, especially concerning the potential reasons why a product is liked or disliked, as well as the emotional links to sensory characteristics (Civille & Oftedal, 2012). Our research group first used free descriptions of the quality experience related to high-quality material in the context of magazines: both paper and print quality are high, and differences

between different versions are preferential. Later on we used this method of combining free descriptions with standard forms of image-quality estimation to assess the subjective quality of cameras (Nyman, Radun, Leisti, & Vuori, 2005), which we called the Interpretation-Based Quality-estimation method (Nyman et al., 2006; Radun, Virtanen, & Nyman, 2006).

1.5.6 Behavioural and psychophysical registration

Given that the information on which image-quality estimation is based is gathered visually, eye-movement registration is commonly used to assess viewing behaviour related to quality changes. Modern technology makes the measurement of eye movements easy and non-intrusive in a way that does not disturb the viewing and estimation process. However, thus far eye registration tends to be used to estimate the spatial distribution of fixations for compiling objective image metrics (Alers et al., 2015; Liu & Heynderickx, 2011) or to develop spatially more precise compression (Kortum & Geisler, 1996; Wang & Bovik, 2001). Eye-movement recordings give information about attention allocation in an image-quality-estimation task, thereby enhancing knowledge about the tasks (Alers et al., 2015; Liu & Heynderickx, 2011) or the different contents (Liu et al., 2013). Chapter 1.4.3 gives more information about eye movements in estimations of image-quality.

Other measures such as brain scanning have also been tried. Electroencephalography (EEG) has been used in relation to JPEG-image compression, responses being observed best in occipital areas (Lindemann & Magnor, 2011). The use of EEG is more common in the estimation of video quality, to see whether changes in quality are visible in the EEG, for instance (e.g. Arndt, Radun, Antons, & Möller, 2014; Scholler et al., 2012). However, these are all outside the scope of this thesis.

2 Research questions and hypotheses

The aim of this thesis is to enhance understanding of the quality-estimation process, especially with naïve participants and high-quality images. The leading questions concern what naïve participants really estimate and how they do it when they are not directed to assess certain aspects of quality. In the following I describe the research questions and the hypotheses in more detail. After each question I indicate which of the studies comprising this thesis address it.

(i) Do naïve participants use the interpretations of the meaning of image features as a basis for quality estimation when assessing the overall quality of an image? Current standards of image-quality measurement (e.g. ISO 20462-1, 2005; ISO 20462-2, 2005; ISO 20462-3, 2012) cannot shed light on this question in that they concern either general quality estimations or ratings of certain image-quality attributes. A combination of standard image-quality methodology and qualitative analysis that is often used in sensory evaluation (Meilgaard et al., 1999) could provide an answer.

Hypothesis 1: Naïve participants base their image-quality estimations on interpretations of the meaning of the image features rather than only on the perceived features.

➔ Studies 1, 2 and 3

(ii) Do the instructions commonly used in image-quality estimation influence viewing strategies even in the case of very similar magnitude-estimation tasks? Various studies attest to the influence of tasks on eye-movement strategies (Castelhano et al., 2009; Mills et al., 2011; Yarbus, 1967), the requirement being that the tasks should differ sufficiently (Andrews & Coppola, 1999). We investigated tasks involving difference and quality estimation, which are very similar magnitude-estimation tasks and are commonly used in image-quality estimation. To our knowledge this was the first time that differences in viewing behaviour have been reported in

two magnitude-estimation tasks based on identical material. We noticed in Studies 1 and 2 that quality could be estimated on two different levels concentrating on the image features or on their interpretation. We posited that the quality estimation would encourage the taking into account of interpretation, thereby directing attention towards semantically meaningful areas. We further posited that the semantically significant areas of the image would not be so important in the estimation of differences, and that the saliency of the area would have a bigger influence on fixation allocation.

Hypothesis 2: Participants engaged in a quality-estimation task concentrate more on semantically important image areas than those engaged in a difference-estimation task.

➔ Study 3 and partly in Study 4

Hypothesis 3: The salient areas are more important in the difference-estimation task than in the quality-estimation task.

➔ Study 3

(iii) What individual differences arise in the rules applied in the image-quality-estimation task? Image-quality estimation tends to be a preference task when high quality is involved. Preferences are subjective, which means there are individual differences. It has been suggested in studies on decision-making that individual differences are attributable to different task construals (Stanovich & West, 2000) or deduction rules (Kruglanski & Gigerenzer, 2011). In the case of quality estimation, the rules can be assimilated from the principles on which people base their estimations (Studies 1 and 2). We therefore examined individual differences in estimation rules, positing that in tasks involving image-quality estimation they would relate to different levels of abstraction, in other words to whether the participants only estimate the changes in image features or also include how such changes influence their interpretation of the images.

Hypothesis 4: Individual differences in how people carry out image-quality tasks are related to the level of abstraction they use in their estimations.

➔ Study 4 (partly in Studies 1 and 2)

(iv) Can eye-movement strategies identify participants applying different estimation rules? Different tasks require different eye movements (Castelhano et al., 2009; Mills et al., 2011; Yarbus, 1967), and individual differences have been reported in studies on decision-making and on eye movements. Participants prefer certain types of viewing strategy even if the task (Boot, Becic, & Kramer, 2009; Rayner, Li, Williams, Cave, & Well, 2007) or the material (Castelhano & Henderson, 2008) changes. In the case of estimation, different estimation rules indicate different task construals. We therefore tested for an association between the use of estimation rules and eye movements, given that such rules relate to understanding the task differently.

Hypothesis 5: Eye movements can reveal differences in estimation rules.

➔ Studies 3 and 4

3 Methods

3.1 Participants

The participants were recruited mainly through the University of Helsinki's students' email lists, and were naïve in relation to the study objectives (Table 2). In all the studies they indicated that they were not involved in photography or image processing professionally, or as a professional-like hobby, and thus were considered naïve in relation to image-quality estimation. Their vision was assessed as normal or corrected-to-normal for near visual acuity, near contrast vision and colour vision (Farnsworth D-15). Most of them were university students, and received cinema tickets or study credits in recompense for their participation.

Table 2. The number of subjects in each study and the numbers of female participants

Study	N of subjects	Females
Study 1	30	17
Study 2	61	46
Study 3	16	10
Study 4: Experiment 1	30	20
Study 4: Experiment 2	30	21
Total	167	114

3.2 Viewing conditions

The images were shown as printed photographs for Studies 1 and 2. The studies were conducted in a room with mid-grey curtains and tablecloths, and adequate lighting. In the case of Studies 3 and 4 the images were presented on computer displays, viewed in a darkened mid-grey room with dim lighting. The distance from the display varied from 80 cm (Study 3 and Study 4: Experiment 1) to 88 cm (Study 4: Experiment 2). At these viewing distances the sizes of the displays varied from 26x20 to 36x23 degrees of visual angle. The viewing distance was controlled with a chinrest only in the experiments related to Study 4. A more detailed description of the viewing conditions is given in the original articles.

3.3 Eye tracking

In Studies 3 and 4 the participants' eye movements were registered while they were viewing the images. For this a standalone eye tracker Tobii x120 (Tobii Technology, Stockholm, Sweden) was used in Study 3 and in Experiment 1 of Study 4. A five-point calibration procedure was applied in these studies. Tobii x120 has a refresh rate of 120 Hz and an accuracy of 0.5 degrees, and two consecutive data points were calculated as being in the same fixation if they were within a 35-pixel (visual angle of 0.67 deg.) radius of one another. We used a free-standing eye tracker (Eyelink 1000 plus) in Experiment 2 of Study 4, with a recording speed of 1000 Hz and an average accuracy of 0.33 degrees. A nine-point calibration was applied at the beginning of the experiment, and drift checks were made between the different parts. The setting used for parsing samples into the fixations and saccades was the threshold velocity of 30°/s and an acceleration of 8000°/s².

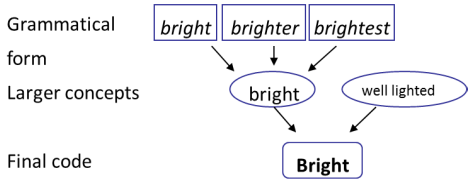


Figure 4. The qualitative coding process, in which synonyms and different forms of the word are combined under the same code

3.4 Qualitative analysis

For the purposes of qualitative analysis (Studies 1, 2 and 4) the participants' explanations were coded according to the principles of grounded theory, the coding starting from the data and larger concepts being gradually formed (Strauss & Corbin, 1998). The codes were formulated so that words referring to the same concept were combined (Figure 4). When the whole data set had been covered the codes were combined into bigger classes, from which the largest groups were selected for the analysis and those with just a few quotations were

left out. Atlas.ti software (Berlin, Germany) (Versions 5 – 7.1.5 depending on the study) was used in the analyses.

3.5 Quantitative analysis

Below I briefly describe why certain methods were used in the studies, and the studies in which they were used are indicated in brackets.

Repeated analysis of variance (rANOVA) (Studies 1 and 4) is suitable for repeated measurements of normally distributed data.

Generalised linear models (GLMs) (Study 3) can deal with data that does not fulfil the requirements of normality by using link function that defines the relationship between the systematic component of the data and the outcome variable (Gill, 2001). This type of analysis was used in Study 3 to examine the differences between the spatial distributions of the fixations between two groups.

Generalised estimating equations (GEEs) (Studies 3 and 4) were used when the data was not normally distributed and there were dependencies attributable to repeated estimations from different participants. They are suitable in the case of non-normal distribution and when the data have missing values, in that they use within-cluster similarity of the residuals to estimate the correlation and thus to re-estimate the regression parameters and calculate standard errors (Hanley, 2003). It is possible to select the distribution that fits the data. GEEs were used in Study 3 to describe the differences in eye movements between the task groups, and in Study 4 to describe such differences between the viewing-behaviour groups.

Correspondence Analysis (CA) (Study 2) shows the relationship between two or more categorical variables in a spatial map, where items frequently occurring together are placed close and variables not occurring together far away. It produces a scatter plot from categorical data, which is a representation of data as a set of points with respect to two perpendicular coordinate axes (Greenacre, 2007). Here we used CA with a Euclidean distance measure and a principal normalisation method, which is suitable when the interest is in the differences between the categories rather than between the variables. Given that different participants gave different numbers of descriptions per picture, we weighted the

final codes so that the sum of descriptions for one image from one participant was equal to one.

Hierarchical cluster analysis (Studies 1, 2 and 4) descriptively classifies cases with similar values on different variables together with creating smaller groups from variables using responses from a set of cases (DiStefano & Mindrila, 2013). Classifications of eye-movement characteristics were used in Study 4 to identify different viewing-behaviour groups, whereas the similarities in images and image attributes were examined in Studies 1 and 2, respectively.

3.6 Eye-movement data analysis

Only the fixations that were inside the image areas were included in the analysis. The first fixation was defined as the first to start after the image appeared on a display. The last ones were excluded given the chance that they might be related to other things than evaluation: in experiments in which the participants themselves stop the viewing by pressing a button, for example, it could be related to preparation for the movement (Kaller, Rahm, Bolkenius, & Unterrainer, 2009). Fixations lasting less than 90 ms or more than 2000 ms were also removed from the data as outliers (Castelhano & Heaven, 2010). The saccade amplitudes were calculated in visual angles using Euclidean distance.

To define the areas to be fixated on we formed a fixation-distribution map of each image convolved with a Gaussian kernel. The full width at half maximum (FWHM) of the Gaussian kernel that defined the size of the patch was set to a visual angle of two degrees (104 pixels in Study 3 and Study 4 Experiment 1, and 146 pixels in Study 4 Experiment 2):

$$\text{FWHM} = \text{visual angle of 2 deg}/2\sqrt{2\ln 2}.$$

Each fixation was weighted according to its duration, and the Gaussian filter approximated the area of accurate vision. In other words, the Gaussian filter was calculated with the standard MATLAB® (MathWorks Inc., Massachusetts, USA) function fspecial, where the FWHM was the standard deviation and the size of fixation was its duration. From this fixation density map (FDM) we defined the regions where the concentration of fixations was high. This calculation of areas fixated on was also used for determining the semantically important image areas.

4 Experiments and results

4.1 Study 1: Can naïve participants say on what they base their quality estimations?

The aim of Study 1 was to enhance understanding of the process via which participants make their estimations. Specifically, we wanted to know whether naïve participants were able to say on what they based them if they were not given a list of terms or training beforehand, and whether they were consistent in their estimations when they used their own words. Standards of image-quality estimation focus on arriving at a single choice or numerical value on the scale of general quality or of some predefined attribute (ISO 20462-1, 2005; ISO 20462-2, 2005; ITU-R BT.500-13, 2012). We wanted to extend the standard methodology by incorporating into the general requirements of psychophysical experimentation a qualitative approach, which is often adopted in sensory evaluation (Meilgaard et al., 1999).

The question we addressed by means of this combined methodology concerned the extent to which people base their decisions on similar rules when estimating changes in sharpness regardless of the image content. We selected sharpness as the variable because it is an important attribute of lens performance, and because it "(1) is readily varied by image processing; (2) is correlated with MTF (Modulation Transfer Function), which can be quantified by measurements from standard targets; (3) exhibits relatively low variability between different participants and scenes; and (4) has a strong effect on image-quality in many practical imaging systems" (Keelan & Urabe, 2004). We wanted to examine the relationship between liking and sharpness ratings with different image contents and lens-like sharpness changes. We were also interested in the extent to which the descriptions concerning the basis of the estimations explained the liking ratings. Our hypothesis was that naïve participants would respond sensibly and consistently with each other when describing on what they based their quality estimations if they could use their own language. We also posited that they would base preference estimations on different interpretations depending on the image content and the level of degradation.

4.1.1 Stimuli

The selected contents comprised five natural images. Four of them were from ISO (ISO 12640-1, 1997) test images denoted as “girl,” “cafeteria,” “fruit” and “bottles”, and the fifth, denoted as “countryside”, was an outside view with green grass, blue sky and forest in the background with a red-coloured bridge in the middle.

Sharpness in all the images was manipulated at the centre (three levels), and as a gradient from the centre to the periphery (five levels) as follows: the optical modulation transfer function (MTF) was used to mimic the sharpness deduction of typical camera lenses. Figure 5 shows the MTF values for the different centre-sharpness groups at 20 lp/mm. These groups could be compared to camera lenses, group 1 representing high quality, group 2 medium quality and group 3 low quality. Fifteen images of each content were presented (3 quality groups and 5 levels of quality).

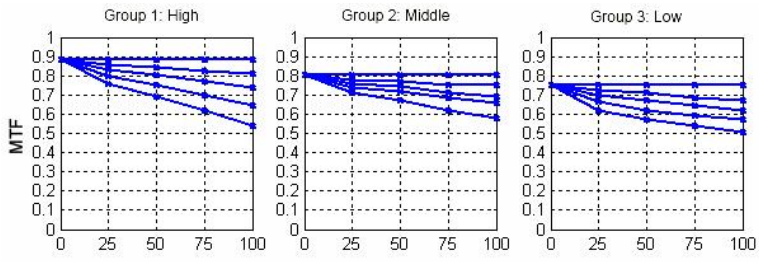


Figure. 5. The MTF values for the different centre-sharpness groups at 20 lp/mm. The X-Axis marks the lens field: 0% marking the centre of the image and 100% the corner.

4.1.2 Procedure

The study was conducted in two stages. In the first stage the participants carried out free-sorting and interview tasks, and then a sharpness-estimation task. Images from one content at a time were randomly placed on a table in the first free-sorting task. The participants were asked to classify these images into groups according to the differences they perceived in them. They were instructed to form at least two and at most fourteen groups, the recommendation being not to produce too many. They were informed that the study was about image-quality, but they were not told what the changing variable in the images was. For each

group the participants gave a preference rating and a general rule they used in their classification (hereafter called a classification rule). Having done the classification they were interviewed and asked to say on what they based it, and what impression they had of the group compared with other groups.

The second stage comprised a sharpness-estimation task requiring the participants to estimated sharpness on an 11-point scale (0 = poor, 5 = moderate, and 10 = good sharpness). They were instructed to estimate the sharpness of the whole picture area. As a reference they were shown an image representing a sharp image (10) and an image that was not sharp (0) from each content.

4.1.3 Results

The participants perceived the changes in the sharpness of the images, the sharpness of both the centre ($F(1,40)=245$, $p<0.001$) and the periphery ($F(2,58)=275$, $p<0.001$) influencing the sharpness ratings. In addition, the contents influenced how the sharpness was perceived (the interaction of the contents and the sharpness from the centre to the periphery $F(9,250)=5.55$, $p<0.001$; and of the contents and the centre sharpness $F(5,136)=4.25$, $p<0.01$). Hence, sharpness degradations were visible, and as expected differently visible, in the different contents.

The participants also indicated how much they liked the images. We examined the association between liking and the sharpness estimations for different contents. When examining the averages per image we noticed that the association between the detection of sharpness and preference differed depending on the content. This is visible in Figure 6 in the angle of the regression lines: the decrease in sharpness in the contents “cafeteria” and “bottles” is clearly considered disturbing (an angle of 0.5 or more), whereas the association is more modest in the other contents. The implication is that even though changes in sharpness can be detected, they do not always influence preference estimations. This was also the case with the general classification rules: for the most part the estimations were based on sharpness (86.2% of all groups). However, the use of this general classification rule also depended on the content: it was applied to only 67.4 per cent of the classifications of the content “girl”. Therefore, the contents influenced which classification rule was chosen. Our aim in this study was to find out which

rules are used if sharpness is not the rule. To this end, we examined in more detail the descriptions of the image groups collected in the interviews.

The interview data was transformed into codes as described in Chapter 3.4. To ensure that the coding was understandable to others and not just to the coder we tested the reliability in terms of inter-coder agreement. A second person coded part of the data and the level of agreement between the two coders was evaluated by calculating Cohen's kappa for each description. Cohen's kappa takes into account the number of codes that would be the same based on chance alone, an informal rule-of-thumb being to regard kappas of less than 0.7 with some concern (Bakeman & Gottman, 1986). However, there is a classification in which kappas below 0.40 are considered poor, between 0.40 and 0.59 fair, between 0.60 and 0.74 good, and between 0.75 and 1.00 excellent (Cicchetti, 1994). In this study, only the code good/pleasant to watch did not reach the limit of fair reliability, and in general the reliability was above good (Table 3).

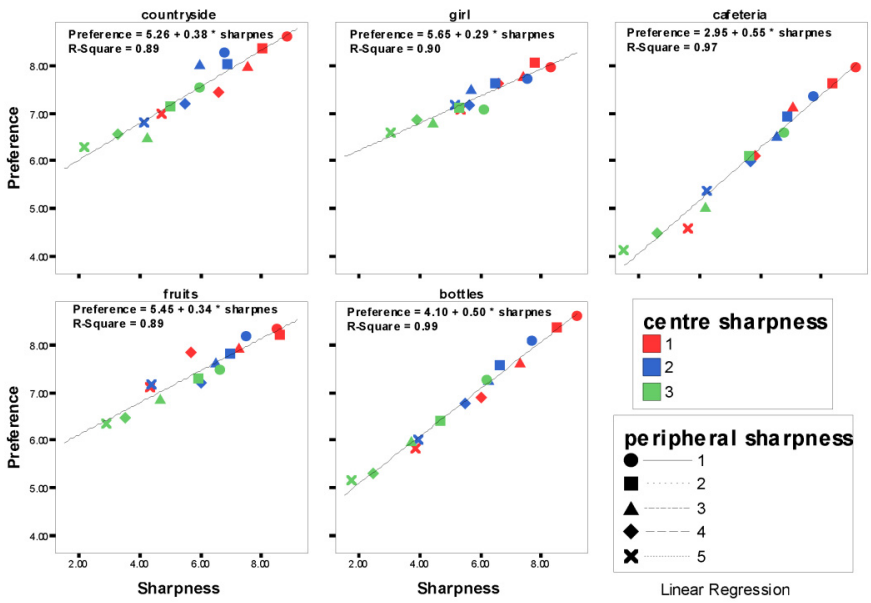


Figure 6. The relationship between sharpness and the preference estimations per image content: the relationship is not the same for all image contents

Table 3. The inter-rater agreement shows how well the two coders arrived at the same taxonomy from the interview material. Inter-rater coding was done for ten interviews. According to Cohen's Kappa, a value of 0 means that inter-rater agreement is the same as would be derived from chance alone and 1 implies perfect agreement.

Descriptions	Cohen's Kappa	
Bright/sunny	1	Excellent
Not sharp	0.865	
Artistic	0.850	
Real	0.838	
Not shiny/dirty/not fresh	0.831	
Sharp	0.808	
Shiny/clean/fresh	0.778	
Professional	0.688	Good
Not alive	0.685	
Amateurish/bad	0.669	
Soft	0.661	
Unreal	0.651	
Light colours	0.588	Fair
Dark	0.532	
Alive	0.517	
Irritating/unpleasant to watch	0.510	
Good/pleasant to watch	0.344	Poor

The free descriptions of the classification basis also differed according to the content (Table 4). The busy images "cafeteria" and "bottles" were influenced the most by the sharpness changes (Figure 6). "Cafeteria" was "irritating to watch" or "bright and sunny" whereas "bottles" looked "shiny and fresh" or "dirty and not shiny". These two contents focused on man-made objects. Fewer such objects were in the contents "countryside" and "fruit", and the descriptions were related to how real the images looked. Interestingly, the image "fruit" started to look artistic when the sharpness clearly decreased. The portrait was estimated as either "professional" or "not alive". The preference ratings in this content were the least affected by the changes in sharpness, probably due to the degradation strengthening towards the periphery and the faces being in the middle.

Table 4. The extent to which sharpness manipulations influenced subjective interpretations varied according to the content. The first column indicates the total number of descriptions and their distribution per image content is presented thereafter (in row percentages). If the descriptions were used equally in all the contents they should be equally distributed (20% per content). However, if a description is more or less important for certain content, the descriptions are distributed proportionally differently. Percentages above 30 and below 10 are emphasised to clarify the between-content differences.

	Total	Contents					Total
		Country-	Girl	Cafeteria		Bottles	
descriptions	Counts	side (%)	(%)	(%)	Fruit (%)	(%)	(%)
not sharp	256	19.9	16.0	24.6	16.8	22.7	100
Sharp	147	20.4	16.3	22.4	18.4	22.4	100
amateurish/bad	112	22.3	17.0	29.5	13.4	17.9	100
good/pleasant to watch	106	19.8	15.1	17.0	18.9	29.2	100
irritating/unpleasant to watch**	90	15.6	**8.9**	**34.4**	14.4	26.7	100
not shiny/dirty/not fresh**	75	14.7	10.7	18.7	18.7	**37.3**	100
shiny/clean/fresh***	64	12.5	**7.8**	15.6	23.4	**40.6**	100
alive	57	17.5	17.5	26.3	21.1	17.5	100
unreal	54	27.8	13.0	13.0	29.6	16.7	100
real*	49	**30.6**	12.2	**8.2**	**30.6**	18.4	100
light colours**	43	14.0	**34.9**	11.6	27.9	11.6	100
not alive*	35	14.3	**37.1**	28.6	**5.7**	14.3	100
dark	28	21.4	14.3	25.0	**7.1**	**32.1**	100
bright/sunny	26	26.9	**7.7**	**38.5**	**7.7**	19.2	100
artistic*	22	27.3	**4.5**	**9.1**	**40.9**	18.2	100
soft	21	19.0	19.0	14.3	28.6	19.0	100
professional	20	15.0	**35.0**	15.0	10.0	25.0	100
Total	1205	19.7	15.8	22.2	18.7	23.7	100

The contents have significantly different amounts of description (X^2 significant on the levels *0.05, **0.01, ***0.001)

The connection between attributes collected from the free descriptions and the preference and sharpness ratings was examined to find out why sharpness in some contents was not assessed as disturbing even if it was visible. The average preference and sharpness ratings related to the same image as the attribute were calculated for each attribute. All the attributes were placed on scales of preference

and sharpness to see when the preference ratings did not follow the sharpness ratings (Figure 7). This examination revealed that there was usually a clear link between sharpness and preference, the descriptions forming attribute pairs such as “pleasant/unpleasant to watch,” “professional/amateurish,” and “sharp/not sharp”. However, there were also attributes that were clearly different from the others, such as “artistic,” “soft” and “light colours”. These attributes were connected with the pictures in which sharpness was perceived as low, but the participants still liked them more than the pictures in which the lack of sharpness created negative impressions (e.g., irritating or dirty). These kinds of aesthetic or stylistic impressions can change the interpretation of a picture completely, after which image fidelity can no longer explain the related preferences.

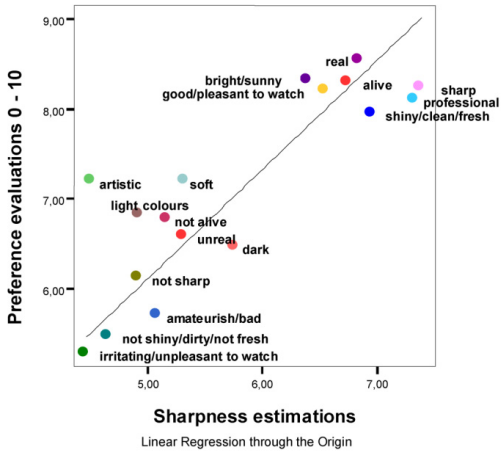


Figure 7. The relationship between the attributes and both the preference and the sharpness ratings presented in a scatterplot

Hence, even naïve participants with no training in image-quality estimation were able to say on what they based their estimations, and were consistent. They also based their evaluations on different interpretations depending on the interaction between the image features and the content. We refer to attributes based on interpretations of the meaning of image features in a certain content as *abstract attributes*, and to those based on the visibility of image features as *feature-based attributes*. We termed the estimation method, which combines

qualitative and quantitative approaches, the Interpretation-Based Quality (IBQ) method. This approach yields additional information on quality estimation from the end-user's perspective.

4.2 Study 2: How non-trained estimators characterise the dimensions of image-quality?

The comparison of imaging devices in terms of quality is an important aspect of product development. The performance of such devices or their components is assessed against the quality of the images they produce. A special characteristic of this kind of quality estimation is the presentation of images with unknown multivariate changes in quality. For this reason it is recommended that the evaluators should be end-users, in other words naïve to the changes in image-quality, primarily because end-users do not based their quality judgments on the technological variables or the physical image parameters, but on what they see – in other words the attributes of the image (Engeldrum, 2004b). However, it is known from the research on multivariate changes in image-quality that the relation between the changes is not directly additive unless they are small (Keelan, 2002). It would therefore serve the purpose of device development also to gather other information to complement the general quality MOS and shed light on the quality experience of end-users. We applied the IBQ method to investigate the rules on which naïve participants base their quality estimations of images with multivariate differences. The main questions addressed in the study concerned the extent to which naïve participants could articulate their rules for quality estimation in a consistent manner, and how far this information could be used to enhance understanding of quality differences in imaging devices.

4.2.1 Stimuli

The stimuli comprised 17 natural image contents. Fifteen of them represented typical home-photography material to show different aspects of image-quality as well as different photo taking conditions. The two remaining contents comprised studio test images taken in two different lighting conditions (D65 light source,

1000 lux and Halogen light source, 10 lux), both of which were designed for the purpose of testing image-quality, especially with regard to camera performance.

The aim was to test different image signal processor (ISP) pipelines, which are used to process the raw image when a photograph is taken (Ramanath, Snyder, Yoo, & Drew, 2005; Zhou & Glotzbach, 2007). ISP operations include colour filter array demosaicking, white balancing, noise filtering, sharpening and colour correction (Bianco, Bruna, Naccari, & Schettini, 2013; Kao, Wang, Chen, & Lin, 2006). This processing allows more natural changes than simply altering a single image feature, the kind of changes that could occur when using a different camera. First we took the RAW output of the image contents captured with a 1.3 megapixel mobile phone camera and ran the ISPs afterwards to simulate the changes produced by a set of processor pipelines.

Study 2 was conducted in two stages, which we refer to here as Part A and Part B. Part A had six pipelines and part B eight. Thirteen different pipelines were tested altogether, one being the same in both parts. The images were presented as 10 x 13 cm paper photographs printed on glossy printing paper. Part A included a total of 102 different test images (17 contents and 6 ISP pipelines), in addition to which was one practice content at the beginning, and two contents were presented twice to check the consistency of the participants' answers. Part B included 103 different test images (17 contents and 8 ISP pipelines). The two sets of images were divided into two and each set was randomised for each participant: half of them saw one image set first while the other half viewed the other set.

4.2.2 Procedure

There were two tasks: ranking and description. First, the participants were asked to rank the images of one image content according to their overall quality, grade 0 being assigned to the one with the lowest quality and grade 10 to the highest. They were instructed to place the other images in between these two so that the distance in quality grades between them was in accordance with the distance in overall quality. The participants were then asked to describe, in their own words, the most important quality aspects of each image.

4.2.3 Results

Sixty-one participants in two experiments gave quality estimations and descriptions for 17 different contents and 13 different camera ISP pipelines. Attributes that were used more than 40 times to describe the reasons for quality estimations resulted in 6,910 quotations distributed over the 20 most frequently mentioned attributes. The association between the attributes and the ISP pipelines were examined to see whether the participants used these attributes in a consistent manner.

CA was used to test the relationship between the attributes and the ISP pipelines, and to identify quality dimensions from the attribute data. CA uses frequencies to plot attributes often occurring together in close proximity, and those not occurring together far apart. This gives an estimate of the performance space in the ISP pipelines. The three-dimensional solution explained 89.0 per cent of the explained variance, the third dimension accounting for 17.7 per cent (inertia 0.165): the fourth, which only explained only 4.4 per cent, was excluded from further examination (Figure 8). The first dimension was named "colour shift" and related to the naturalness of images and the overall colouring, hence the white balance settings (Figure 8). The second dimension was called "darkness", even though the other end of the dimension was "graininess" (Figure 8a). We attributed this to the camera's sensor gain, which is increased to deal with dark targets, hence reducing darkness but causing graininess. The third dimension was called "sharpness" (Figure 8b). This constitutes the subjective quality space for ISP pipelines, which is examined below.

Figure 9 shows the subjective quality space and the distribution of the ISP pipelines in it. The pipelines are marked to show whether the general quality is high, medium or low. In addition, on each pipeline is information including its number and the sub-study it was from, as well as the general-quality average. The first dimension of the subjective quality space was "colour shift", which was related to the naturalness of colours, and differentiated the high-quality pipelines from the others (Figures 8 and 9). Therefore, the high-quality pipelines were different from the others in their natural colours and the lack of colour shift. The other dimensions, "darkness" and "sharpness", distinguished the attributes related to medium and lower quality in terms of pipeline performance. This

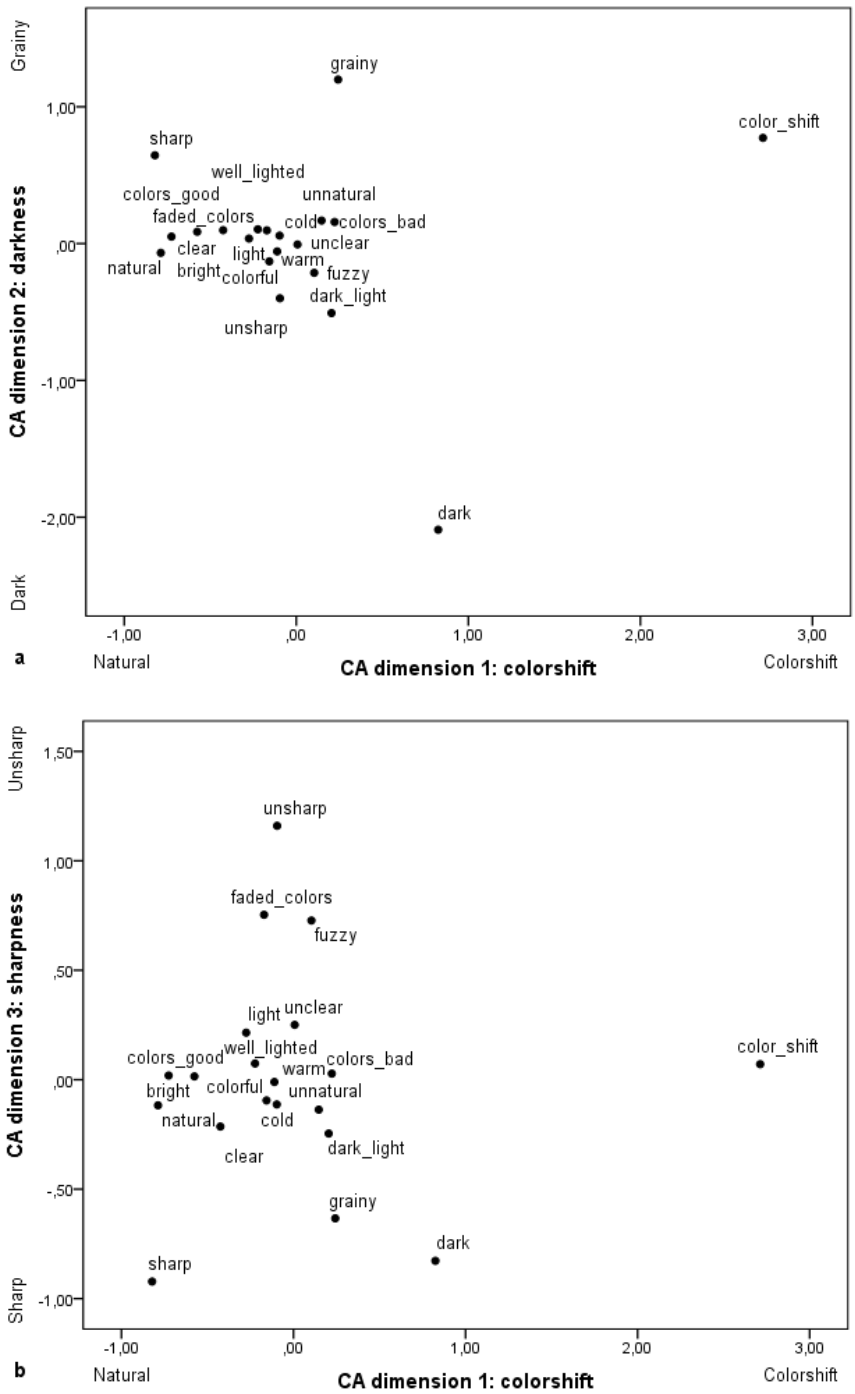


Figure 8. The CA scatterplots show how the attributes are distributed in the subjective quality space. 8a presents dimensions 1 and 2, and 8b presents dimensions 1 and 3.

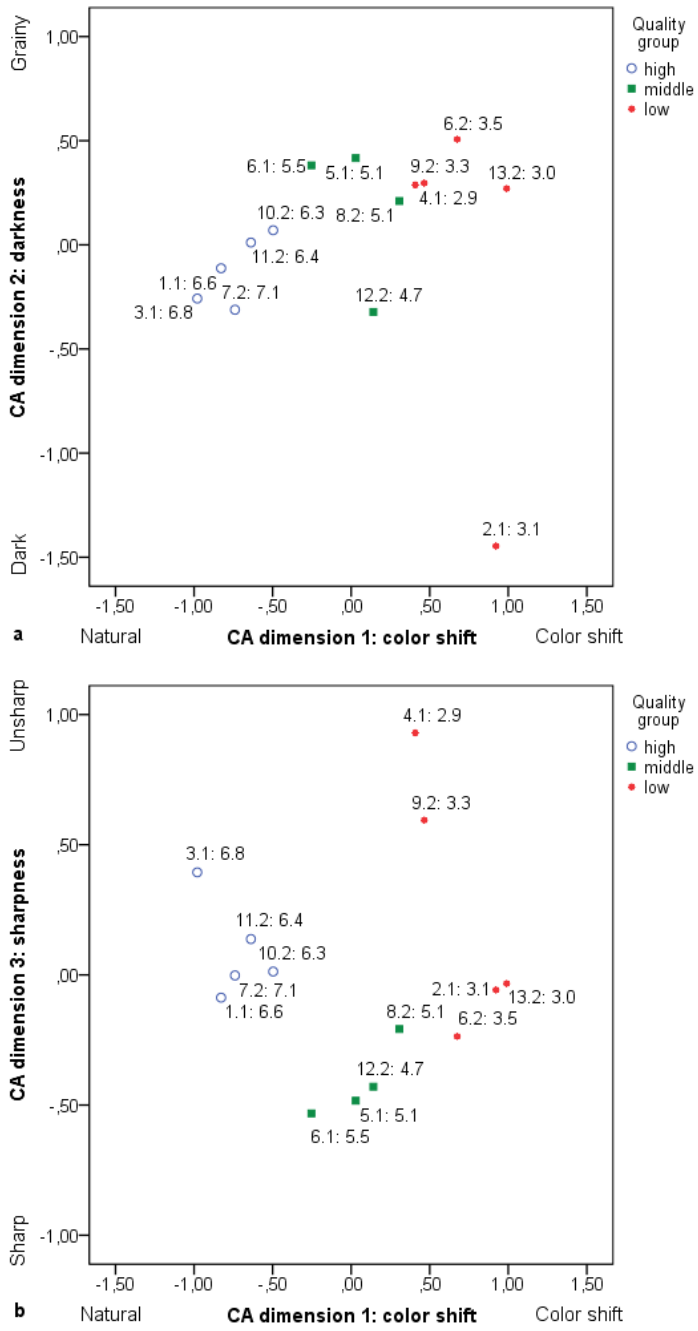


Figure 9. The CA scatterplots show how the ISP pipelines are distributed in the subjective quality space. The numbers on each pipeline denote the number of the pipeline, and the study and the average MOS. Figure 9a presents dimensions 1 and 2, and 9b presents dimensions 1 and 3.

analysis shows that descriptions from naïve participants can be used in conjunction with more traditional estimations (in this case MOS) to enhance understanding of the device's or the component's performance.

Therefore, according to the results of Study 2, IBQ estimation methodology makes it possible to use naïve participants' descriptions to form a subjective quality space. It also allows for a more detailed description of quality than when the examination is limited to the averages of overall quality estimations. Furthermore, it shows that even though quality estimations are subjective, people base them on similar rules.

4.3 Study 3: Do small changes in instructions change the way people seek information from images?

To enhance understanding of the process of image-quality estimation, we examined the strategies applied in two tasks that are commonly used for this purpose: the estimation of quality and the estimation of difference. Both are magnitude-estimation tasks based on the same material, and only the instruction changes. We measured the strategy by means of eye-tracking, which is a non-intrusive and objective measure of behaviour suitable for tasks in which the information search is visual. Differences in viewing strategies have been noted in tasks that are somewhat different, such as active and passive viewing tasks (Andrews & Coppola, 1999), as well as memory and visual-search tasks (Castelhano et al., 2009). Viewing behaviour has also been examined in tasks that include free viewing and quality estimation (Alers et al., 2015; Larson et al., 2008; Liu & Heynderickx, 2011; Vu et al., 2008). The focus in free-viewing tasks is on the most prominent regions of interest, whereas image-quality estimation involves wider scanning for quality clues (Alers et al., 2015): comparisons of tasks with global and local degradations have revealed differences in fixation allocations only if the degradations were local (Vu et al., 2008).

Here we examined strategy differences in two magnitude-estimation tasks. We posited that the quality-estimation task would resemble a preference task when high-quality images were shown, whereas difference estimation would be a detection task. Even though the differences in the instructions are small, the

information requirements direct the information search differently depending on the image areas, given that the cognitive relevance of the areas depends on the instruction (Henderson et al., 2009). We also posited that the semantically meaningful areas of an image would be more influential in quality estimation, which is a preference task, than in difference estimation, and that difference estimation would require more searching. We expected the quality estimations to concentrate on semantically meaningful areas, which would be contrary to the results reported in Alers et al. (2015). However, our material comprised high-quality images and we expected the difference estimation to reflect cases in which the artefacts are clearly visible. We further posited that the salient areas of the images would be more important in difference estimation, because the semantic content should not matter in the detection task and attention could be focused more strongly on salient areas.

4.3.1 Stimuli

The stimuli comprised seven image contents representing everyday photographing. We used the photospace that indicates the distance and illumination distribution of a large sample of photographs as a selection guide (I3A, 2007; Keelan, 2002). The material included close-ups of people (2), people further away with many surrounding details (2), a town scene with people (1), a town scene without people (1) and a nature scene (1). All the other contents were normal everyday photographs, with the exception of two. One of these, which was developed specifically for the evaluation of colour still image processes, showed a woman with many objects around her (Salmi, Halonen, Leisti, Oittinen, & Saarelma, 2009). The other was a town scene with people sitting in an outside cafeteria, which is ISO-recommended content for evaluating the results of image processing (ISO 12640-1, 1997).

We selected both structural (blur, noise, jpeg-compression) and non-structural (white point, and increased and decreased luminance) manipulations to give us a wide variety. We also divided the contents into two groups to increase the number of different manipulations, each group including a close-up of people, people further away and scene images without people. These groups were processed differently. The three contents were subjected to five different types of processing:

blur, noise, the white point, and increased and decreased luminance. For four image contents, JPEG2000 compressions were made with the publicly available codec Kakadu 6.0 (www.kakadusoftware.com), using three different bitrates: 0.1068, 0.21173, and 1.708 bpp.

4.3.2 Procedure

The experiment consisted of two parts: a memory task, and difference or quality-estimation tasks. The participants' vision was tested when they came into the laboratory. Then they were given instructions and the eye tracker was calibrated. Their eye movements were recorded when the test images were visible on the display. The first instruction was for the memory task: the participants were asked to view each image and when it had disappeared to write down what was in it as if explaining it to another person. They saw one image from each of the test-image contents in this phase of the memory task. Having completed the task the participants were divided into two groups, one was given the quality-estimation task and the other the difference-estimation task. Both groups were shown the original image first, then the manipulated version with the same content, and then the original image again. Between the showings there was a fixation point on a mid-grey background. The setting was modified from a previous study examining perceptual differences (To, Lovell, Troscianko, & Tolhurst, 2010). The participants in our study were able to look at the images for as long as they wished. Figure 10 depicts the procedure as a flow chart.

Members of the difference-estimation group were instructed to estimate the size of the change in an image pair (referred to hereafter as the difference task), and members of the quality-estimation group to estimate the extent of the change in quality (referred to as the quality task). The estimation scales were based on a reference image pair, which was shown at the beginning of the task, after four showings of practice contents, and throughout the test as every tenth image pair. The participants were informed that in numerical terms the amount of change or of change in quality in the reference pair was 20, and that the value 0 indicated no visible difference between the two images in question. The reference-image content depicted a parking lot, which was processed using different ISP pipelines creating optical artefacts. These images therefore simultaneously showed

moderate changes in colour as well as in sharpness and graininess. The reference images also showed a moderate change in quality in multiple image artefacts. For this purpose, the value 0 was defined as no visible difference or no visible difference in quality, and the value 20 was defined only in terms of the image pair because graphical scales with quality terms associated with different steps cannot be divided into intervals of equal size (Teunissen, 1996). The test-image pairs were presented in five different random orders, so that four participants (two from each task) always did the test with the same randomisations.

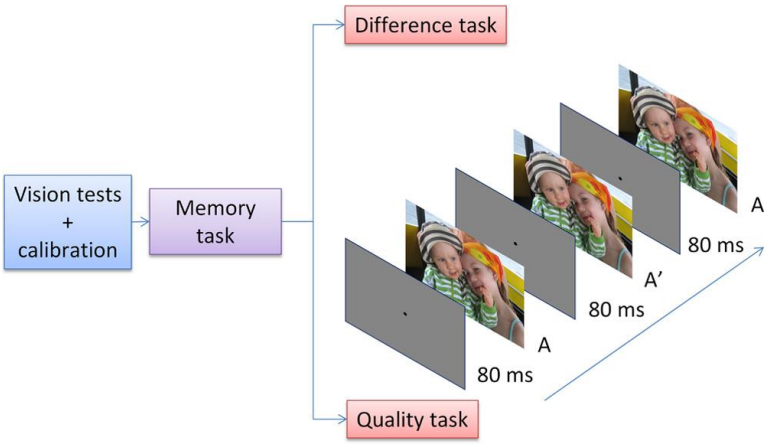


Figure 10. The procedure presented as a flow chart. A is the original image and A' is the manipulated image of the same content. The participants decided how long they would look at each image.

4.3.3 Analyses

4.3.3.1 Defining semantic regions of interest (ROIs) and areas fixated

The semantic ROIs were defined from the eye movements during the memory task. The fixated areas were estimated from a fixation distribution map, as described in Chapter 3.6. The cut-off point for the z-axis of the resulted fixation density map (FDM) was 0.25, which our qualitative examination of the distributions showed to be the value that best suited the most images. The areas fixated were calculated for each image across the observers in a similar manner from both quality and difference-estimation tasks, with a cut-off point of 0.02.

4.3.3.2 *Defining salient areas*

The low-level salient areas were defined using Walter and Koch's (2006) Saliency Toolbox 2.2 (http://www.saliencytoolbox.net/ downloaded in July 2011). It is based on the modelling work of Itti and Koch (2000) and calculates the saliency map for the image using information on contrast, colour and orientation weighted with the winner-take-all maps. We used the toolbox's default settings and because our images contained humans we added a skin-colour feature with the weight of one. The areas with positive values were considered salient.

4.3.4 Results

There were no differences between the groups in the estimations of magnitude the participants gave for the images (Wald $\chi^2(1)=1.0$, $p>0.05$). However, there were differences in viewing strategies (Table 5 presents the main results). The participants engaged in the difference-estimation task looked at the images for a longer time (Wald $\chi^2(1)=9.2$, $p<0.01$) and needed more fixations (Wald $\chi^2(1)=7.2$, $p<0.05$) than those doing the quality-estimation task. The average fixation durations per image per participant did not differ according to the task (Wald $\chi^2(1)=1.1$, $p>0.05$), but there was an interaction between the task and the content (Wald $\chi^2(5)=19.8$, $p<0.01$) as well as between the task and the manipulation (Wald $\chi^2(6)=13.6$, $p<0.01$). This indicates that the influence of a task becomes visible in fixation duration only if the type of test material is taken into account. We further examined the duration of the first fixations, which are used to plan subsequent eye movements throughout the scene (Castelhano & Henderson, 2007) and in this case when the content is known they yield information about the processing of the first actively chosen fixation point (Holmqvist et al., 2011). The first fixations were longer in the quality task than in the difference task (Wald $\chi^2(1)=7.5$, $p<0.01$), meaning that planning where to look next took longer in the former than in the latter. The average saccades amplitude per image was also longer in the difference task (Wald $\chi^2(1)=8.7$, $p<0.01$), and the task-content interaction was significant (Wald $\chi^2(5)=15.4$, $p<0.01$). Therefore, the fixations were further apart in the difference task than in the quality task, and there was less detailed examination of one area with repeated fixations.

Table 5. The medians of the variables describing strategy in the quality and difference tasks, as well as the significance of the between-task comparisons

	Quality task	Difference task	p-value
Viewing time (ms)	2465	3914	**
Fixation duration (ms)	333	319	ns
Saccade amplitude (deg)	5.4	6.1	**
First fixation duration (ms)	350	300	**
Fixation count	6	10	**
Area fixated on (%)	15.2	26.1	***
Proportion in salient areas (%)	14.0	15.0	*
Proportion in ROIs (%)	75.4	64.8	***

ns = non-significant, *=p<0.5, **= p<0.01, ***=p<0.001

We also examined the spatial distribution of the fixations. First, a larger area was fixated on in the difference task than in the quality task (Wald $\chi^2(1)=232.2$, $p<0.001$). The medians of the areas fixated on were 26.1 and 15.2 per cent in the difference and quality tasks, respectively. The fixations in the quality task covered 16.7 per cent of the image area on average, and in comparisons between the two tasks, only 4.1 per cent of the area fixated on in the quality-estimation task was not fixated on also in the difference task. It seems that those engaged in the difference-estimation task fixated on the same important areas as those engaged in the quality task, but these areas were not enough: a further large area was needed to cover the search in the difference task. Next we analysed what kinds of areas were fixated on in these different tasks.

To define the types of areas in the different image contents we calculated the salient areas from the low-level image features as well as the semantic regions of interest (ROIs) from the eye movements recorded in the memory task. In these analyses we estimated the group-level differences in the areas fixated on. The salient areas were widely distributed across the images, and the areas considered salient differed depending on the contents. In the content “woman”, which depicts a woman sitting by a table with many different objects around her, only 6.4 per cent of the image area was considered salient. The corresponding figure for the content “scenery”, which depicts a nature scene with water, forest, rocks and sky, was 10.4 per cent, the largest proportion of all the contents. Similarly,

the size of the semantic ROIs depended on the content: the semantic ROIs covered 6.1 per cent of the image area in the portrait of a boy, compared with 41 per cent in a busy image of an outside cafeteria. Therefore, the areas considered salient did not vary in size according to the image contents as much as the areas considered semantically important.

The time spent looking at these areas differed between the tasks. A higher proportion of time was spent looking at semantic ROIs in the quality task than in the difference task (Wald $\chi^2(1)=251.0$, $p<0.001$), and vice versa in the salient areas (Wald $\chi^2(1)=4.3$, $p<0.05$) (Figure 11). The proportion of time spent fixated on a certain area also depended on the interaction between the task and the contents (semantic ROIs: Wald $\chi^2(5)=118.2$, $p<0.001$; salient areas: Wald $\chi^2(5)=52.6$, $p<0.001$). The biggest between-task differences in attention allocation concerned contents with strong attention attracters, such as faces or large areas considered semantically important (the content "cafeteria"). The information from the strong attention attracters seemed to be enough in the quality-estimation task, whereas attention was also actively allocated outside this area in the difference task. It therefore seems that semantically important image areas are more important in the quality-estimation than in the difference task. Such areas are fixated on in the latter task, as is a large area in addition.

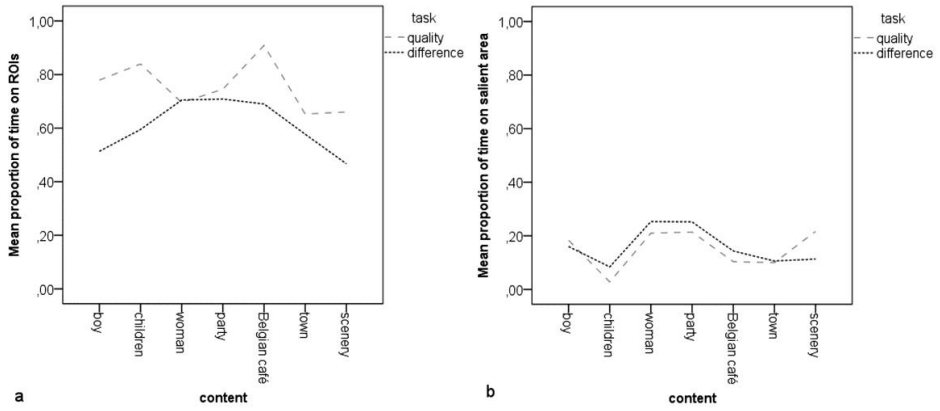


Figure 11. The average proportions of time spent on semantic ROIs (a) and salient areas (b) of all the time spent looking at images per image content.

To shed further light on the relationship between salient areas and semantic ROIs we examined their common relationship and the areas fixated on. The salient areas that were fixated on were often also within the semantic ROIs (Figure 12), the proportions falling outside being only 1.7 per cent in the quality task and 3.5 per cent in the difference task. The area that was both salient and in the semantic ROIs comprised less than five per cent of the whole image area on average, but it nevertheless accounted for 13.8 per cent of the fixations in the quality task and 12.0 per cent in the difference task (Figure 12a). It thus seems that the salient areas of an image are important only if they are also semantically important. This supports the notion that saliency models work, because most objects are salient (Einhäuser, Spain, et al., 2008).

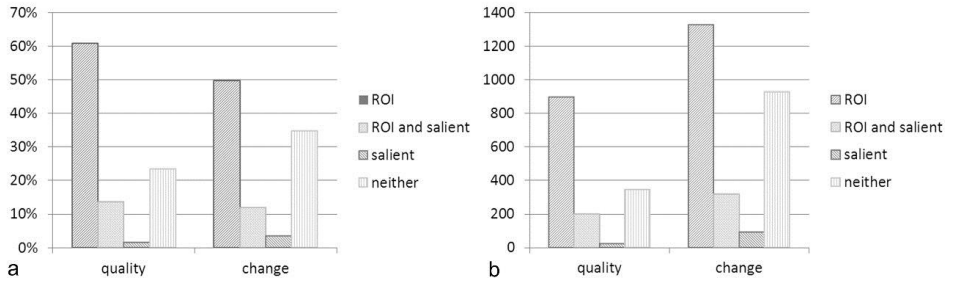


Figure 12. The proportions (a) and the numbers (b) of fixations on different image areas for the two tasks

The results show that a small change in the instructions influences viewing behaviour, even when comparing magnitude-estimation tasks. The viewing times were shorter, and the viewing concentrated more on the semantically important areas in the quality-estimation than in the difference-estimation task. The semantically important areas were attended to in the difference task as well, but this information was not sufficient and other large areas were also fixated on. The salient areas were important if they were also semantically important.

4.4 Study 4: Are individual differences in viewing behavior related to different estimation rules in a quality-estimation task?

The findings from Studies 1 and 2 revealed that, depending on the material, people use different decision-making rules when estimating quality. It was shown in Study 3 that the instructions influence viewing behaviour even in two quite similar magnitude-estimation tasks. In the case of high-quality images quality estimation tends to be a preference task and therefore subjective. Subjectivity means that there are individual differences, which have been linked to the use of different deduction rules (Kruglanski & Gigerenzer, 2011). We posited that quality estimation in the case of high-quality images could be a task in which subjective differences are related to different deduction rules. We further assumed that estimation rules on which quality estimations are based could be used to access deduction rules. Here, we used here the term *estimation rules* to refer to the set of attributes on which people base their estimations.

Individual differences in eye movements have also been reported (Andrews & Coppola, 1999; Boot et al., 2009; Castelhano & Henderson, 2008; Rayner et al., 2007). We wanted further to find out whether such individual differences are related to differences in estimation rules. We conducted two experiments to examine this relationship. In the first one we used the IBQ method, to elicit estimation rules: the method allows people to describe freely on what they base their preference estimations. The aim of second experiment was to confirm the results of the first using a larger image set and a different method of measuring the estimation rules. The participants estimated how important the different attributes collected in Experiment 1 were for their preference estimation. Both experiments involved the recording of eye movements and analysis of the relationship between individual viewing tendencies and estimation rules.

4.4.1 Experiment 1: Stimuli

As stimuli we used eight different image contents representing normal home photography, selected to show different everyday photographic content in line with the guidelines for photospace examination (I3A, 2007; Keelan, 2002). The

images featured included close-ups and long shots of people, and some contained no people.

Different ISP pipelines (see Chapter 4.2.1 for a more detailed explanation) were selected to show the small differences that create different impressions of the images. We selected images that had received equal general-quality MOSs in previous image-quality tests, therefore their general performance was similar enough and the differences in general perceived quality were minor. Furthermore, image-quality specialists selected the test images and pipelines so as to ensure that the image processing would not change the visibility of details or make the images appear unnatural, thereby meeting the image-quality requirements of discriminability and identifiability (Janssen & Blommaert, 2000). The purpose of this was to ensure that the differences in the images were related to subjective preferences, and not to a lack of compliance with basic image-quality requirements.

4.4.2 Experiment 1: Procedure

The experimental procedure followed the IBQ method with magnitude estimation. After the vision tests the participants were instructed to rate how much they liked the following images on a scale from one to ten (1 not at all pleasant, 10 very pleasant), after having looked at them, and after each rating briefly to write down their reasons. They were encouraged also to include their impressions among their reasons. The test leader first went through three practice images with different content than the test images to make sure that the participants had understood the instructions correctly. The images were shown in random order and the participants themselves decided for how long they would look at them. A fixation point appeared on the stimulus display on a middle grey background for about 80 ms before each image appeared. The fixation point was shown in the four corners and in the middle, to diminish the central bias in image viewing (Tatler, 2007).

4.4.3 Experiment 1: Analyses

4.4.3.1 Defining the fixated areas

The fixated areas were estimated from a fixation-distribution map as described in Chapter 3.6. We determined the magnitude of the area one participant fixated on in one viewing. The cut-off point for the z-axis of the resulting fixation-density map (FDM) was 0.001, which our qualitative examination of the distributions showed was the value that best suited most images.

4.4.4 Experiment 1: Results

The participants were classified into viewing-behaviour groups by means of hierarchical cluster analysis to facilitate examination of individual differences in viewing behaviour. We included the average values of fixation duration and saccade amplitude, as well as the area fixated on for each content and participant. These eye-movement measures divided the participants into three viewing-behaviour groups that differed in fixation duration ($F(2,27)=91.7$, $p<0.001$) (Table 6). The differences in other eye-movement measures were not significant (saccade amplitudes: $F(2,27)=2.7$, $p=0.09$; fixation counts: $F(2,27)=1.5$, $p=0.25$; area fixated on: $F(2,27)=1.8$, $p=0.19$), therefore the classification was based on fixation duration. The groups were named the short, medium and long fixation-duration groups.

Table 6. Viewing-behaviour measures among the three groups. The averages, standard errors of the means and standard deviations are presented for the different variables. The significance (sig.) shows whether or not the groups differed from each other in the rANOVA.

	Group 1=short N=9			**Group 2=medium N=16**			**Group 3=long N=5**			**sig.**
	Mean	SE of Mean	SD	Mean	SE of Mean	SD	Mean	SE of Mean	SD	
Fixation duration	252.0	2.2	31.5	302.6	2.2	42.6	356.4	4.9	54.0	***
Saccade amplitude	5.6	0.09	1.34	5.1	0.06	1.2	5.1	0.13	1.37	ns
Fixation count	48.2	2.5	36.5	54.4	1.6	31.6	36.8	2.4	26.5	ns
Area fixated (%)	27.0	1	11	31.7	1	13	24.4	1	11	ns

***= $p<0.001$, ns=$p>0.05$

The final aim of the study was to find out whether individual differences in viewing behaviour were related to different estimation rules. We coded the reasons for different preference estimations as described in Chapter 3.4. Table 7 shows the most frequently used attributes. The last column indicates whether or not the attribute was considered feature-based, meaning based on the visibility of the image features, or abstract, meaning based on an interpretation of the features in a certain image content.

Table 7. The frequencies of subjective attributes and whether an attribute was considered feature-based or abstract

Subjective attributes	Frequency	Feature-based (F) vs. Abstract (A)
bright	50	F
warm	56	A
atmosphere good	59	A
colours natural	63	A
colours bad	81	F
dark	89	F
light good	117	F
grainy	136	F
sharp	185	F
not sharp	206	F
colours good	241	F

To facilitate examination of the general estimation rules we classified the attributes, combining those that related to the same concept. For example, all the attributes related to sharpness were placed in the same class regardless of whether they were related to comments on sharpness or blur. Table 8 shows the largest attribute classes and their frequencies. Four classes with frequencies of above 300 were used for further analysis given that the frequency of the subsequent class was considerably lower (graininess with 150 quotations). The classes chosen for further examination were sharpness, colour, illumination and abstractness. Sharpness included comments related to sharpness or fuzziness, or the visibility of details. Colour contained all the descriptions related to colour

unless they were related to a higher concept such as natural colours, in which case they were classified as abstract attributes according to the concept of naturalness. Abstractness included concepts requiring more elaboration and interpretation, such as dirty, calm or fresh. The descriptions in the illumination group were connected with light, brightness and darkness, and those in the graininess group included comments on whether or not an image looked grainy.

Table 8. The subjective attribute classes: the frequencies indicate how many times the attributes belonging to these classes were mentioned

attribute class	frequency
sharpness	477
colour	444
abstractness	419
illumination	346
graininess	150

The contents influenced which rules were used as a basis for the evaluations in all the attribute classes (Sharpness: Wald $\chi2(7)=30.4$, $p<0.001$, Illumination: Wald $\chi2(7)=30.9$, $p<0.001$, Colour: Wald $\chi2(7)=29.1$, $p<0.001$), except the abstractness group (Wald $\chi2(7)=13.8$, $p=0.055$). Therefore, different attributes were important in different contents. However, the processing did not influence which attributes were used as a basis for the estimations (Abstractness: Wald $\chi2(2)=1.2$, $p>0.05$, Sharpness: Wald $\chi2(2)=0.3$, $p>0.05$, Illumination: Wald $\chi2(2)=5.5$, $p>0.05$, Colour: Wald $\chi2(2)=3.7$, $p>0.05$). This could be attributable to the ISP pipelines used, which process images taken under different circumstances differently in that, as expected, the use of attributes correlated with the objectively measured changes in images (see the original article for more details). Therefore, the image content determined the classification rules.

Even though the image content influenced which attribute classes were used in the estimations, the viewing-behaviour groups differed only in the use of abstractness (Wald $\chi2(2)=10.0$, $p=0.007$) (Sharpness: Wald $\chi2(2)=1.9$, $p=0.37$, Illumination: Wald $\chi2(2)=1.2$, $p=0.55$, Colour: Wald $\chi2(2)=0.0$, $p=0.99$). In other words, the viewing-behaviour groups differed in terms of whether the participants also based their estimations on abstract attributes or only mainly on

feature-based attributes. When we examined the viewing-behaviour groups we noticed that the group viewing images with fixations of medium duration used the most abstract attributes (Figure 13), and that the group viewing images with long fixations seemed to use the most feature-based attributes. It also seems that assessments of images with humans in them are always based on abstract attributes (Figure 13). It may be that perceiving humans, and especially faces, always involves interpretation, and it has been shown that basic facial expressions are rapidly identified and categorised (Palermo & Rhodes, 2007).

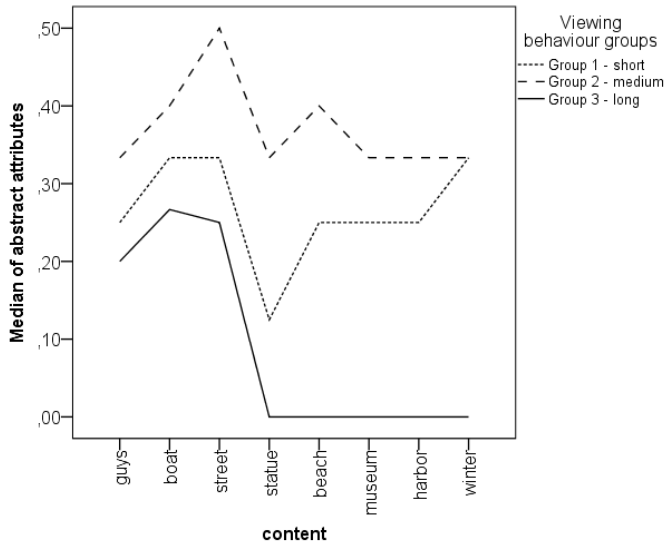


Figure 13. The median proportions of abstract attributes of all image attributes mentioned presented in terms of the different contents and the viewing-strategy groups.

The results show that people displaying different viewing behaviour base their quality estimations on different rules. These rules are related to the emphasis they place on feature-based attributes vs. abstract interpretations of changes in meaning. It seems that when the images have strong attention catchers such as human faces some interpretation is always included. The viewing-behaviour groups using the most features-based attributes in their estimations viewed images with the longest fixations. It may be that the examination of image features needs more information from one fixation to facilitate discrimination of the low-level features in one place instead of taking in the whole image with its

meanings, where several shorter fixations are needed. The group with medium-duration fixations based their estimations on the most abstract attributes.

4.4.5 Experiment 2: Introduction

The results of Experiment 1 were interesting in that the use of abstract attributes was not related to the viewing-behaviour group with the longest or shortest fixations, but to the one with medium durations. We wanted to examine this result further in a new experiment.

Because Experiment 1 had just eight image contents, and the contents had a clear influence on how the quality was estimated, we decided to use a larger set of images. Second, the fact that the researcher constructed the attribute classes for the measurement of estimation rules means that there may have been some bias in how the attributes were classified. Therefore, this time the participants rated the importance of different attributes related to image-quality for their preference estimations. The selected quality attributes were the ones most frequently used in Experiment 1, which are also in line with previous studies on image attributes (Pedersen, Bonnier, Hardeberg, & Albregtsen, 2010), and in Studies 1 and 2 in this thesis. As in Experiment 1, the importance of the attributes was rated after the participants had estimated the general quality of the images in a similar task, because estimating certain aspects of image-quality can direct attention in a specific way. Third, to specify the areas the participants attended to when concentrating on feature-based or abstract estimation rules the semantically important image areas were calculated from the eye movements in a memory task in which the instruction was to recall what was in the image. In sum, the aim in Experiment 2 was to examine individual differences in the quality-estimation task with a larger set of images and a different method of estimating the importance of estimation rules than in Experiment 1.

4.4.6 Experiment 2: Stimuli

Experiment 2 had 24 content items: 10 representing normal consumer photography and 14 images from the LIVE multiply distorted image-quality database (Jayaraman, Mittal, Moorthy, & Bovik, 2012) with known changes in images. The consumer-photography contents included six used in Experiment 1,

excluding the portrait images (“guys” and “boat”) that seemed to be estimated differently from the rest, always including an interpretation of the image content (see Figure 13). The four additional contents depicted a restaurant garden, a bar at night with neon lights, a candle, and a pet rabbit in a cage. These images were processed with six different ISP pipelines, including those used in Experiment 1. The home-photography contents were chosen to show multivariate quality changes, whereas the LIVE images were selected to show known quality changes. The “baby girl” image from the LIVE multiply distorted image-quality database was excluded because it was a portrait. The smallest distortions of each content on the variables blur, jpeg, noise, blur and jpeg, and blur and noise, as well as the reference images without any processing, were chosen from the LIVE images.

4.4.7 Experiment 2: Procedure

The experiment had three parts: a memory part, a free-estimation part and a scale-estimation part. The procedure followed the common guidelines as in Study 1. First, the participants’ vision was tested. Next they were given the general instructions concerning the study and then the eye tracker was calibrated. The participants were told that this was an image-quality study, that we were interested in their opinions and that there were no wrong answers. They were instructed to sit in front of a stimulus display in front of which was the eye tracker, and to place their chin on the chin and headrest. A nine-point calibration was used. After this came the memory part: the participants were told that they would see 24 image contents and they would need to remember what was in the images later. At this point the reference images from the LIVE database and the images of one pipeline were shown in a random order. Each image was shown for five seconds.

The scale-estimation part of the experiment consisted of the task from Study 1. The participants were asked to rate how much they liked this specific version of the image content on a scale ranging from one to nine (1 = not at all pleasant, 9 = very pleasant), and after each rating to briefly give their reasons in writing. In all cases the rating screen appeared after they had been looking at the image for eight seconds. They were encouraged to describe the good and bad features in the image in accordance with their own impressions. The experiment leader first

went through three practice images with different content than the test images to make sure that the participant had understood the instructions correctly. At this stage each participant saw one version of the 24 image contents, selected at random from all six versions of each one, and the chosen images were shown in a random order.

The third part comprised the scale estimation. The participants' task was to estimate how much each feature influenced their preference rating of that specific image on a scale ranging from one to nine (1= not at all, 9 = a lot). The viewing time was eight seconds, after which the answer screen appeared. The features used were taken from the most frequently used attributes in Experiment 1: sharpness, illumination, colours, graininess, atmosphere and naturalness. The question was worded thus: Estimate how much _____ influences your liking of the image. For each scale the participants estimated one version of each image content. The versions were again randomly selected for each scale, so that each version was seen once (6 different versions and 6 scales) in the scale part. The scales as well as the contents within them were always presented in a random order to each participant.

There was a fixation cross in the middle of the screen on a mid-grey background for 500 ms before each image appeared. The drift was corrected before the second and third stages of the experiment, as well as between each scale.

4.4.8 Experiment 2: Analyses

4.4.8.1 Defining semantic ROIs

The semantic ROIs were defined in accordance with the eye movements in the memory task. The fixated areas were estimated from a fixation-distribution map, as described in Chapter 3.6. The cut-off point for the z-axis of the resulting fixation-density map (FDM) was 0.15, which our qualitative examination of the distributions showed to be the value that best suited the most images. The areas fixated on were calculated for each image content across all participants.

4.4.9 Experiment 2: Results

The participants were again classified into viewing-behaviour groups by means of hierarchical cluster analysis (see Chapter 3.5). The eye movements recorded in part 2 when the participants were estimating general image preference, as in Experiment 1, were included. The variables included the average fixation duration and saccade amplitude among the participants, and the proportion of fixations that were on the semantic ROIs was also used. Semantic ROIs were used to measure the spatial distribution of the fixations, given that in Experiment 1 we could not find any differences between the viewing-behaviour groups in the areas fixated on. Furthermore, we posited that interpretations of the meaning of image features, i.e. the abstract assessment of images, would need more fixations on the semantic ROIs than when the focus was on feature-based attributes.

The hierarchical cluster analysis classified the participants in three viewing-behaviour groups. The rANOVA that was conducted to investigate the differences between the groups again showed that they varied in fixation duration ($F(2,24)=107.2, p<0.001$) (Table 9), as in Experiment 1. The fixation counts were also different ($F(2,24)=37.8$, $p<0.001$), because unlike in Experiment 1 the viewing time was fixed. The groups did not differ in saccade amplitude ($F(2,24)=0.8$, $p>0.05$), and the fixations were similarly distributed in the semantic ROIs ($F(2,24)=0.2$, $p>0.05$). The classification therefore resembled the one derived in Experiment 1, in which fixation duration was the only differentiating factor in the viewing-behaviour groups. It thus seems that fixation duration is the factor showing individual tendencies in quality-estimation tasks.

Next we analysed how the participants in these viewing-behaviour groups estimated the importance of different attributes related to image-quality. Such information was obtained in Experiment 1 from the attributes on which the participants said they based their estimations. In Experiment 2, the participants estimated the importance of the attributes on scales when assessing different versions of the same image contents. This estimation was clearly different from those in Experiment 1, and similar results in both experiments with their different methods and different-sized image set would mean a link between viewing behaviour and the estimation rules.

Table 9. Viewing-behaviour measures among the groups: the averages, standard mean errors and standard deviations are presented for the variables related to viewing behaviour, the significance (sig.) showing whether the groups differed from each other in the rANOVA.

	Group 3=short N=7			**Group 1=medium N=11**			**Group 2=long N=9**			**sig.**
	Mean	SE of Mean	SD	Mean	SE of Mean	SD	Mean	SE of Mean	SD	
Fixation duration	249.5	2.3	30.3	298.0	2.1	34.5	347.1	3.2	47.0	***
Saccade amplitude	4.4	0.1	1.2	4.7	0.1	1.1	4.8	0.1	1.3	ns
Fixation count	26.3	0.2	3.1	22.5	0.2	2.5	19.2	0.2	3.0	***
Fixations in ROI (%)	63.0	1.3	17.4	61.9	1.0	16.9	64.0	1.3	19.2	ns

***= $p<0.001$, ns=$p>0.05$

Given that the order of importance of the attributes within a participant was of interest, the participants' ratings were normalised by dividing the ratings they gave for all scales by the mean of one participant's ratings. This reduces the effect of different scale usage and emphasises the order of scales within each participant. GEEs, which take into account the repeated nature of estimations, were used to examine these normalised scores (see Chapter 3.5). The variability of contents was taken into account in examining the groups' importance ratings of the image-quality attributes. In general, the ratings did not differ among the groups (Wald $\chi2(2)=2.5$, $p>0.05$), which was to be expected given that the values were normalised within subjects. However, the viewing-behaviour groups gave different ratings depending on the scale they estimated, since the interaction of group and scale ratings was significant (Wald $\chi2(10)=24.7$, $p<0.01$). Hence, different viewing-behaviour groups assessed different attributes as important in their preference estimations.

The groups differed in their estimations of the attribute atmosphere (Wald $\chi2(2)=11.0$, $p<0.01$), but not in the other scales (naturalness: Wald $\chi2(2)=4.0$, $p>0.05$; graininess: Wald $\chi2(2)=2.2$, $p>0.05$; sharpness: Wald $\chi2(2)=3.3$, $p>0.05$; illumination: Wald $\chi2(2)=0.2$, $p>0.05$; colours: Wald $\chi2(2)=1.2$, $p>0.05$).

Atmosphere could be considered an abstract attribute, and therefore this finding confirms the result from Experiment 1 indicating that the viewing-behaviour group with medium fixation durations used more abstract attributes than the other viewing-behaviour groups. However, these groups did not differ in their estimations of naturalness, another abstract attribute. This may have been because naturalness is connected with the basic image-quality requirement of identifiability and is always taken into account. Feature-based attributes are related to the other image-quality requirement of detectability, and also seem to be taken into account even when abstract interpretations of the meaning are included. Whether or not the interpretation of perceived image change is estimated is more of individual choice and is related to the person's conception of the task.

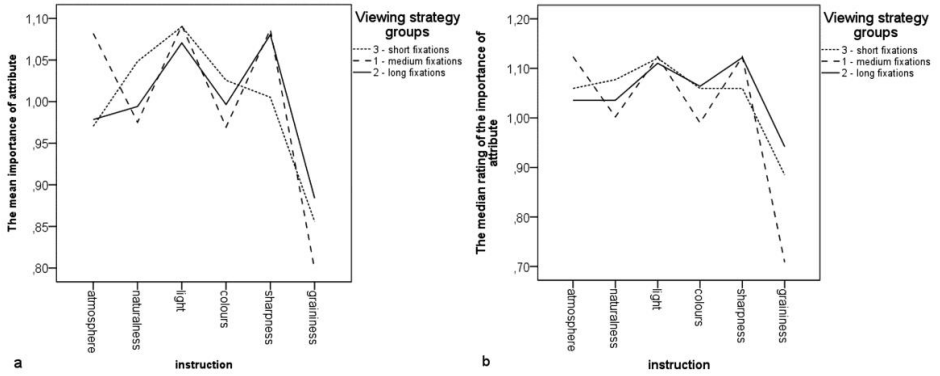


Figure 14. The mean (a) and median (b) importance of the image-quality attributes among the viewing-behaviour groups

Figure 14 shows how the different viewing-behaviour groups estimated the importance of the image attributes. The only difference in importance rating concerned atmosphere: the group with medium fixation durations rated the atmosphere as more important for their preference estimations than the two other groups. This also confirms the results of Experiment 1 indicating that the medium fixation-duration group adopts the most abstract estimation rules.

5 Discussion

The main aim in this thesis was to enhance understanding of image-quality estimation, with specific reference to naïve participants and high-quality material. To shed further light on the subjective quality experience we also refined the standard measurements of image-quality. In this concluding chapter I will first introduce methods that could bring new insights into how image-quality is experienced, and then I will concentrate on the process of quality estimation from the psychological perspective.

5.1 The measurement of image-quality

Given the above-mentioned aim, we developed a method that would reveal more about quality experience among naïve participants than traditional methods, especially with regard to high-quality material. Most of the standard methods of image-quality estimation currently in use are based on psychophysical-measurement traditions and emphasise the assignment of values corresponding to the strength of the perceptions (ISO 20462-1, 2005; ITU-R BT.500-13, 2012). However, such methods either assign a common value to the estimation, or guide participants to concentrate on certain image attributes. Careful thought should be given to the instructions the participants receive. When they are asked to assess sharpness in terms of liking and visibility, for example, the contents determine the relationship, in other words the importance of sharpness for liking (Study 1). In addition, the instruction to estimate changes in quality or changes in general led to different eye-movement behaviours (Study 3). We wanted to know what naïve participants really estimated in images with multivariate changes if their attention was not directed to certain aspects of quality and they estimated quality in general. We considered two methods that could give additional information on quality estimation when combined with the traditional psychophysical approach: the IBQ method and eye-movement measurement.

5.1.1 Interpretation-based quality – the IBQ method

We started using the IBQ method, which combines qualitative and quantitative approaches, in our research on image-quality estimation to find out on what naïve participants base their quality estimations (Nyman et al., 2006, 2005; Radun et al., 2006). This information is especially useful in product development, when the changes in image-quality are multivariate and the aim is to elicit the views of end-users who tend to be naïve with regard to image-quality estimation. In addition, it makes it possible to examine naïve observers' estimation rules, thereby providing new information on the process of quality estimation.

The strength of the IBQ method over others is that it does not direct observers' attention to specific predefined aspects of image-quality. For example, attribute scales direct participants' attention only to certain attributes of the image. Furthermore, the language used might be difficult to understand, especially among those with little experience of image-quality estimation and the use of scales. Using scales may also be challenging when the image-quality is high given the potentially small contribution of separate image attributes to quality, the degradation in quality coming from the discrepancy between the attributes and the content. In such cases, image-quality estimation could be considered a task in which people estimate what is suitable for each image content in a certain situation.

The aim in Studies 1 and 2 was to find out whether naïve participants were able freely to describe the basis of their quality estimations, and whether they produced consistent results as a group even without any training. Our conclusion was that when they used their own vocabulary they were able to explain on what they based their estimations in a consistent manner. This was also the case in Study 2 in which we defined image-quality dimensions using this approach. The approach was adopted from the field of sensory evaluation, which, however, often uses descriptions devised by panels with training in the vocabulary and the rating process rather than naïve participants (Meilgaard et al., 1999). Nevertheless, it seems that naïve participants have the vocabulary to talk about the basis of their estimations of image-quality. This may be because we live in a world that strongly emphasises visual information. Furthermore, people might base their estimations on how they interpret the meaning of the image features, in which

case they do not need any specific vocabulary to describe changes in images in a rather systematic manner.

The IBQ method also has its limitations. It is suitable for high-quality material with multivariate changes in particular. However, if the quality range is wide it may produce obvious answers that could be estimated from the images by means of objective measurement or just by viewing the image. The method is also somewhat time-consuming and quite heavy on participants: the number of estimations is limited and should concentrate on high-quality material with small quality changes. In addition, the qualitative approach has its requirements: there must be enough subjects and the coding of the data may be laborious. However, when the aim is to elicit the views of naïve participants about their experiences with high-quality material, the method supplies valuable information.

5.1.2 Eye-tracking in image-quality estimation

Another method we used was eye-movement tracking, which I discuss here in relation to the information it provides in estimations of image-quality. However, I should point out that, the focus in the eye-movement tracking carried out in this thesis was on detecting different viewing behaviours and not on material-related differences. On the positive side, eye tracking is an objective method for measuring behaviour and does not intervene in the process of quality estimation because post-calibration the task can be done without much interference. If this objective non-intrusive method could be used to examine differences in behaviour it would make the detection of strategies or even subgroups possible without lengthening the experiments very much, or changing the instructions.

To detect different viewing behaviours in tasks that are commonly used in eye-movement studies we focused on the tasks of quality and difference estimation (Study 3). Earlier studies on eye movements in quality-estimation tasks tended to compare free-viewing and quality-estimation tasks (Alers et al., 2015; Liu & Heynderickx, 2011; Ninassi, Le Meur, Le Callet, Barba, & Tirel, 2006; Vu et al., 2008). However, the instruction was not the only factor that changed in most of these studies in that original images comprised the material in free-viewing tasks as opposed to processed images in quality estimation (Liu & Heynderickx, 2011; Ninassi et al., 2006; Vu et al., 2008). In these task comparisons the fixations

concentrated more strongly on the areas in which local quality degradations were visible in the quality-estimation than in the free-viewing task, whereas no task differences in fixation allocation were detected in the case of global degradations (Vu et al., 2008). It was reported in a recent study comparing these tasks with processed images viewed in both tasks that the quality losses did not consistently modify the allocation of visual attention, and that free viewing in general concentrated more strongly on the most prominent regions of interest whereas scanning was called for in the quality-estimation task (Alers et al., 2015).

Study 3 compared two magnitude-estimation tasks, the material being identical in both, and differences in viewing behaviour were detected. Participants engaged in the quality-estimation task concentrated on semantically important image areas with fewer fixations and a longer first fixation to plan the subsequent fixations, whereas those engaged in the difference-estimation task also scanned a wide area in addition to the semantic regions of interest. However, it was concluded in a study comparing free-viewing and quality-estimation tasks with identical images that the fixations were more widely distributed in the latter than in the former (Alers et al., 2015), possibly because of the differences in material between the two studies: we used high-quality material, whereas the materials Alers et al. (2015) used varied more in quality-levels. This could lead those engaged in the quality-estimation task towards finding the artefacts from the images, reflecting the difference task in our Study 3 in which we found that only a small change in instruction led to differences in viewing behaviour. It is therefore highly important to formulate the instructions given to participants with care when subjective image-quality studies are planned.

We were also able to differentiate viewing-behaviour groups based on individual data from the eye-movement recordings in the quality-estimation task (Study 4). These groups differed in fixation duration, a result that two separate experiments yielded. Furthermore, the three viewing-behaviour groups differed in how much they emphasised interpretations related to the changes in the image (Study 4). It thus seems that eye movements can be used to detect subject groups using different evaluation rules.

5.2 The process of quality estimation

Clarifying the process of quality estimation from a psychological perspective was the main aim of the thesis. In this section I consider this process from three angles: estimation rules, context dependency and subjectivity. Estimation rules refer to the quality requirements on which naïve observers base their estimations, and may enhance understanding of the process. Context dependency concerns the interaction between the instructions, the material, subjective expectations and so forth, all of which influence the process. Subjectivity is considered separately because it is often treated as unwanted noise in the measurements. I treat it as a factor that yields valuable information about preference estimations. Finally in this section I introduce a model of image-quality estimation that facilitates comparison of how the material and the instructions influence the estimation task and its performance.

5.2.1 Estimation rules

Quality requirements are reflected in the rules on which people base their estimations, in other words in the aspects of image-quality to which they pay attention. These rules derive from, among other things, the instructions (Study 3), the material (Studies 1 and 2) and individual preferences (Study 4), and from the overall context in which the estimations are made. The requirements are reflected in viewing behaviour (Studies 3 and 4) as well as in the estimations (Studies 1, 2 and 4).

To clarify how naïve participants use quality attributes we formulated subjective dimensions of image-quality (Study 2) and also examined the general rules applied in quality estimation (Study 4). The quality dimensions referred to the image-quality of camera phones, with multivariate changes from which three quality dimensions emerged: the naturalness of the colours, darkness and sharpness (Study 2). In the case of high-quality images with multivariate changes the most commonly used attribute classes were sharpness, colour, abstractness, illumination and graininess (Study 4).

At about the same time as Study 2 was published, Pedersen et al. (2010) wrote an article defining six image-quality attributes for colour prints based on a

literature review: colour, lightness, contrast, sharpness, artefacts and physical attributes. Colour, lightness and sharpness are the same as in our studies. We also refer to artefacts such as graininess. The participants did not comment on physical-quality attributes related to physical parameters that affect quality, such as paper properties and gloss, probably because this was not a variable that changed in the material. The attributes we discussed did not include contrast either, possibly because of its close relation with sharpness and the fact that it may be less familiar as a concept to non-trained naïve participants. Pedersen et al. (2010) make no mention of abstract attributes such as natural and artistic, which might reflect the impressions and interpretations on which naïve participants base their estimations and are often ignored in studies on image-quality. Nevertheless, it seems from the free descriptions of our naïve participants that they constitute a common basis for quality estimation, especially regarding changes in high-quality images.

Image-quality requirements have been defined as identifiability and discriminability (Janssen & Blommaert, 2000). Easy identifiability comes from naturalness and the prototypical aspects of images that make them easy to recognise. The challenging nature of identifiability was evident in the image-quality dimensions when naturalness was the attribute differentiating the highest-quality ISP pipelines from the other lower-quality pipelines (Study 2). The differences in the lower-quality images were related to feature-based attributes such as sharpness and darkness. Low-level quality features, which relate especially to discriminability, are examined to some extent (Study 4). Higher-abstraction-level attributes such as naturalness in general and colours served to distinguish the high-quality images from the others (Study 2). However, abstract impressions were also used sometimes when quality degradation was clear and the defining factor seemed to be the combination of image content and its features (Study 1). Most participants based their estimations on both abstract and feature-based attributes, although some focused mainly on the latter (Study 4). It thus seems that naïve participants also tend to interpret the meaning of perceived changes and to base their estimations on them (Hypothesis 1). Their estimation rules are thus determined in interaction between the material and the meanings it conveys, and the subjective preferences and expectations of the

participants. These findings imply that participants change the basis of their quality estimations even during the same experimental session, depending on the material, when they are allowed to use their own language as they would if they were looking at the images as end-users.

5.2.2 Context dependency

The context comprises aspects such as the material used, the instructions given and the participants' own expectations and preferences. The material refers to the image contents and processing that are used as stimuli in the research. The estimations are indicative of how well the context and the material fit together when the basic image-quality requirements of discriminability and identifiability are fulfilled.

The influence of image contents and expectations was evident in Study 1, in which the extent to which the participants liked images that were not sharp depended on the content. Another factor that influenced the ratings was a change in sharpness from the centre to the periphery: when there were fewer objects in the periphery the change in sharpness had the smallest effect (contents "fruit" and "girl"). However, this does not explain why the changes in sharpness clearly influenced liking in the content "bottles", for example. The reason why changes in sharpness in some contents did not directly reduce the level of liking was that the participants interpreted the image differently when it became less sharp (Study 1). For example, if the sharpness was reduced enough the bowl of fruit started to look artistic. Perceived sharpness had more of a negative effect on preference when the contents depicted mostly man-made artefacts than in scenes with natural objects or humans in them. It has been observed that contrast degradations matter more with man-made than with natural content (Tinio et al., 2011). It may be that sharp details are not as important in natural objects, and humans have learned to change their interpretations according to what they see. If, for example, a mountain does not look as sharp as usual it is probably due to fog: it does not change the properties of the mountain but it does influence interpretations of the weather.

The influence of the content was also evident in Study 4: the viewing-behaviour group that generally based its estimations only on feature-based attributes also

used some abstract attributes when humans were the main objects in the image. Faces are strong attention catchers and are always fixated on in images (Cerf et al., 2009). There are indications that detecting facial configurations may be obligatory, and that basic facial expressions are rapidly categorised and identified (Palermo & Rhodes, 2007). It seems that there is always some interpretation involved.

Instruction is another contextual factor in that it changes the task and determines what is cognitively relevant in the images. We compared two magnitude-estimation tasks that are commonly used in studies on image-quality (Study 3). The influence of the task on eye movements has been noted in clearly different contexts, such as active and passive viewing tasks (Andrews & Coppola, 1999), tasks involving search and memorisation (Castelhano et al., 2009), and in a set of preference, search, memorisation and free-viewing task (Mills et al., 2011), for example. However, the viewing behaviour was different even in our two magnitude-estimation tasks. In the quality-estimation task, the longer first fixations directed attention to the semantically important regions of the image, whereas the first fixation was shorter in the difference-estimation task but proceeded to a longer examination of larger image areas (Study 3), including the semantically important areas but also others. The semantically important regions were thus viewed for longer in the quality task than in the difference task (Hypothesis 2). The salient image areas appeared more important in the difference task than in the quality-estimation task, but in the main these areas were both semantically important and salient (Study 3). Even though the difference between the tasks in fixation allocation to these salient areas was small it was significant, thereby supporting Hypothesis 3.

Viewing strategies in preference tasks have been examined with respect to the time course of the viewing (Antes, 1974) and in comparison with other tasks, the fixations at the beginning of viewing tasks being shorter than those at the beginning of memory and free-viewing tasks (Mills et al., 2011). We found in Study 3 that the task groups did not differ in fixation duration but they did in the image areas fixated on. It seems that with high-quality material the estimations concentrate on impressions and the interpretation of how well an image fits the

idea the participant has of its purpose. Contextual information includes not only the image content, but also interaction with its usage in general.

We also examined the areas fixated on in the two tasks in relation to their semantic importance and salient areas based on low-level image features. The salient areas fixated on were defined as also being semantically important (Study 3). This attests to the saliency models working mainly because semantically meaningful objects are often salient, which means that they are fixated on because of their meaning, not because of their low-level features (Einhäuser, Spain, et al., 2008). Therefore the fixation in these tasks is not on the salient areas per se, but on the semantically important areas that tend to be salient.

5.2.3 Subjectivity and individual differences

Given that the quality-estimation task is subjective, we expected to observe individual differences in the participants' answers and behaviour. Individual differences were evident in the qualitative examination of the IBQ data in Studies 1, 2 and 4, as well as in eye movements in the two experiments conducted in Study 4. In the context of decision-making, individual differences have been identified in deduction rules (Kruglanski & Gigerenzer, 2011). We also observed individual differences in estimation rules in a quality-estimation task (Study 4), although these related to the level of abstraction and not, for example, to the varying emphasis on low-level image features – meaning that one person might examine images based on sharpness whereas another will concentrate on colours. A different level of abstraction in the responses implies that some people estimate a decrease in sharpness in the image of a human whereas others detect the change in sharpness and interpret it as making the skin of the person in the image look softer. This confirms that participants differ in the level of abstractness of their estimation rules (Hypothesis 4).

Fixation duration was the factor that differentiated the viewing-behaviour groups in both experiments conducted for Study 4. The eye-movement features included in the classification were fixation duration, saccade amplitude and a measure of the spatial distribution of fixations. The spatial measure in Experiment 1 was the area fixated on. It did not influence the classification of individuals, and therefore the proportion of fixations on the semantically

important image areas was used as the spatial measure in Experiment 2. However, again the duration of the fixations was the only factor that classified the observers into different viewing-behaviour groups. Individual differences in fixation duration have been identified previously in different search tasks, for example (Boot et al., 2009), and in comparisons of viewing strategies in the case of images with faces and scenes in them (Castelhano & Henderson, 2008). Furthermore, stability in fixation duration has been reported among specific individuals in different tasks, specifically reading, face processing, scene perception and counting, and two visual-search tasks (Rayner et al., 2007). Our results confirm that fixation duration is also related to individual tendencies in image-quality estimation.

One of our hypotheses was that the viewing-behaviour groups would understand the task differently and would therefore use different estimation rules. The groups did, indeed, differ in their estimation rules. In Experiment 1 of Study 4 they used different levels of abstract attributes, whereas in Experiment 2 of Study 1 the difference was in the estimation of the importance of the atmosphere. The atmosphere is an abstract attribute of image-quality, and therefore the results of Experiment 2 confirm and refine the findings from Experiment 1. In sum, individuals engaged in an image-quality-estimation task differ in their level of abstraction, in other words in whether or not they include interpretation of the observed changes (Hypothesis 4). The difference in estimation rules is also related to differences in viewing behaviour (Hypothesis 5).

In addition, the results of the two experiments conducted in Study 4 similarly showed how the estimation rules were related to the different viewing-behaviour groups: the group with medium fixation durations used more abstract attributes and rated the atmosphere as a more important estimation rule than the others. Therefore, whether an interpretation is included or not is related to the duration of the fixations, and this was shown in both of the experiments.

Why is fixation duration related to the level of abstraction in estimation rules? It has been associated with processing difficulty (Henderson et al., 2013): the longer the fixation the deeper is the processing (Holmqvist et al., 2011). Furthermore, short fixations are common in search tasks, for example (Mills et al., 2011; Rayner, 2009), whereas slightly longer durations are typical in free

scene viewing (Mills et al., 2011; Rayner, 2009) and in the memory task (Mills et al., 2011). We found that the groups with medium fixation duration used the most abstract estimation rules. The duration times in these groups bore most resemblance to those in the free-scene-viewing and memorisation tasks, whereas in the groups with short fixations the durations were most closely related to those in the search tasks. In general, free-viewing and memorisation tasks tend to include some kind of interpretation of the image contents, whereas in search tasks the requirement per fixation is simply to detect whether the target is or is not in that certain place. One explanation could be that the groups with the shortest fixation durations used simple rules so that the processing per fixation was easy. Furthermore, the group with long fixations could have tried to assess many feature-based attributes in each one. Further, individuals estimating images based on impressions need to cover all the semantically important areas, and the assessment is a combination of the general impressions formed at all these fixation points. As a consequence, less information will be retrieved from one fixation point when estimations of quality are based on impressions, in which case they are based on many repeated fixations. Fixation duration depends on how much information can be extracted from one area and how many repeated fixations are needed for this. It has been suggested that repeated fixations on one area indicate the importance of the area more strongly than fixation duration (Castelhano et al., 2009). However, verifying the reasons for the link between fixation duration and estimation rules is beyond the scope of this thesis. A further investigation could involve the examination of temporal aspects of viewing in relation to different estimation rules.

5.2.4 The process of visual high-image-quality estimation

In the following I explain the special characteristics of high-image-quality estimation compared to other types of tasks used in subjective estimations of image-quality (Figure 15). First, I identify the factors that determine how the task is understood (environment), and then I discuss the consequences in terms of understanding and carrying out the task (see Figure 15).

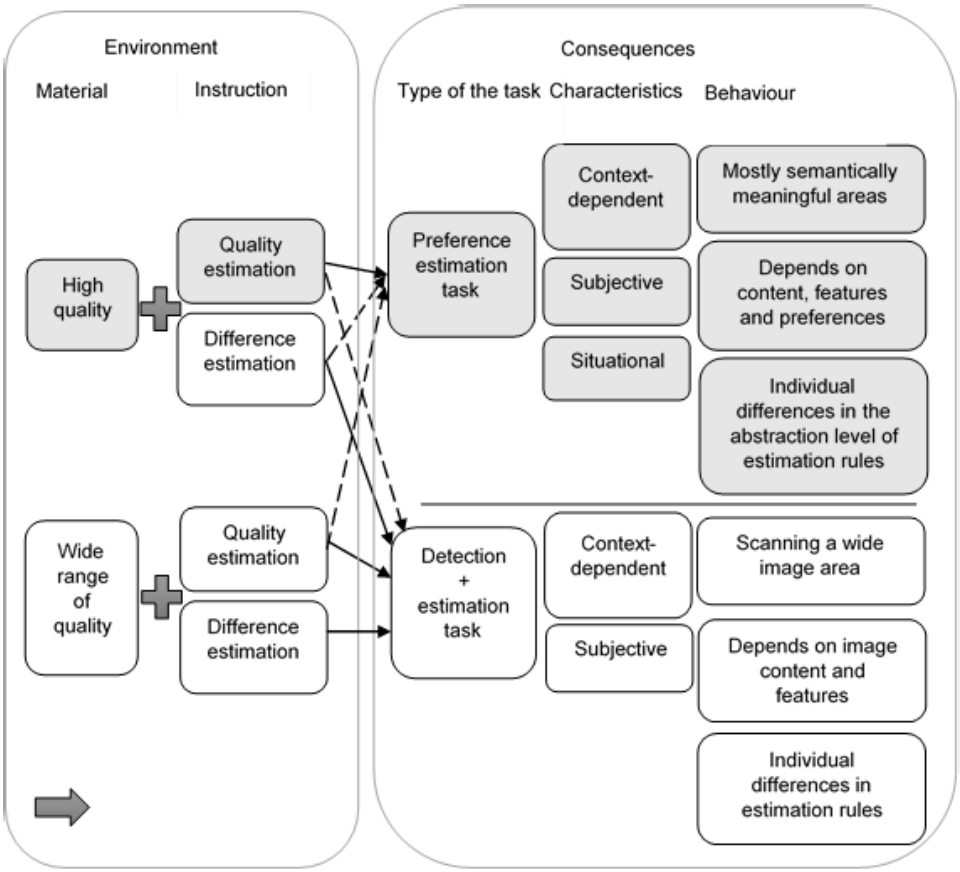


Figure 15. A model of visual-quality estimation that takes into account the material and the task, and their influence on the process. The boxes with a grey background describe the focus of this doctoral dissertation.

The instructions and materials given to participants define how the task is understood. Personal factors such as expertise also influence this, but they are not included in Figure 15 for the sake of simplicity. The material consists of image contents and features, but the focus here is on the range of quality presented. If the range is wide, the changes in some images are highly visible and easily detectible. However, all the basic requirements of image-quality are met in the high-quality range and the mere detection of changes or artefacts is not enough: the answer lies in how well the image contents are presented with certain of its features.

The quality- and difference-estimation tasks are separated in the model, and both are used in studies of image-quality. The arrows in Figure 15 represent the likelihood of a task being carried out as a preference or a detection and estimation task. The solid arrows indicate a high likelihood and the dashed arrow a low likelihood. The focus in the difference-estimation task is on finding changes or artefacts (detection) and estimating their magnitude (estimation). If the material comprises images of very high quality with changes that are barely noticeable it could become a preference task (the dashed line in Figure 15). The other instruction often given to observers is to estimate quality. If the quality range is wide this may well make it a detection task with an estimation component, although with certain combinations of image contents and features it may be a preference-estimation task. An example of this is described in Study 1, when the participants sometimes thought the changes in sharpness improved the image content. However, estimating the quality of high-quality images is likely to be a preference task, especially with naïve observers. In that case, detecting the changes or artefacts is not enough in that their influence on the image is not unambiguous and depends more on personal preferences. Nevertheless, some people still treat this as a detection and evaluation task. One such group could be image-quality experts, who have the training, the vocabulary and clear estimation criteria (the dashed line in Figure 15).

Image estimation is always context-dependent and subjective. The context-dependency is evident, for example, in how the visibility of the features depends on the contents. The subjectivity in detection and estimation tasks is related to the image features that are used as the estimation rule, which tends to concern low-level image attributes and possible artefacts. Visual searching for areas in which different degradations or artefacts are visible also leads to search behaviour involving the scanning of wider image areas than just the semantically meaningful. Estimations of the amount of visible difference are somewhat subjective, although objective measures of image-quality tend to succeed in predicting such estimations given that the change in artefacts is clear. These objective measures are not as effective when the image-quality is high and the changes in artefacts are small.

The preference-estimation task is context-dependent and subjective, but also situational. Objective changes in high-quality images may be minor, and the context dependency of a preference task is related to the content and features of the image, but also to individual preferences based on past experience and situational factors. Here, situational factors are related to the expectations created by the environment and the individual's interpretations of them. Individual interpretations reflect the subjectivity that is evident when image attributes are used as estimation rules. The important attributes may relate to the low-level features of the image, but also to the level of abstraction - meaning how much observers use their interpretations of the changing image features in their estimations. Given that the changes in images tend to be minor, the information search focuses on the semantically meaningful areas. Changes that are clear and extend beyond the semantically meaningful areas will naturally be included in the estimation.

5.3 Limitations

As a limitation of the study, it must be noted that the participants in all the experiments were mainly Finnish-speaking university students, a somewhat young and selected population. Age-group and cultural differences have been reported in some image-quality studies, as well as in studies on eye movements. Older participants were found to prefer somewhat warmer white points than younger ones, although the difference was not large (Beke et al., 2008). Difference have also been found in the eye movements of American and Chinese participants viewing images in that the Chinese spent more time looking at the background information (Rayner et al., 2007). Moreover, some of the interpretations may be culture-related. There are differences, which should be taken into account when the results are interpreted even if they are not expected to be drastic.

Furthermore, all the interviews were performed in Finnish as well as the coding, but the results were reported in English. This might have changed some nuances of the codes, however, we think translating the final codes does not change the findings, because already in the coding phase synonyms are combined

into larger codes. This means that final codes describe a general level, where the importance of small nuances in language has already evaporated.

Furthermore, the number of participants is somewhat limited, especially in the eye tracking studies. In Study 3, we tested twenty participants, but due to problems in the quality of eye movement tracking, we had to exclude four. This resulted in a somewhat limited number of participants, but after examining the data carefully, we concluded that the results reflected group behaviour more than the behaviour of few individuals only and therefore decided not to take more participants. In Experiment 1 of Study 4, we expected the two groups of participants rather than three and estimated 30 participants to be enough. After the somewhat surprising finding of three subgroups, one of which being very small, we noticed that the result was not strong enough. However, rather than taking more participants to Experiment 1, we decided to make another experiment, with a different method and more images. When this experiment showed a similar result as Experiment 1, we thought the result was strong enough to report. However, we do not know why the medium length fixation durations were related to the use of abstract estimation rules. This must be examined in the future studies, by for example examining the viewing behaviour in the tasks of quality and difference-estimation with a larger group of participants and expanding this to the examination of estimation rules.

Another limitation of the studies comes from the material chosen. The number of image contents and manipulations is always a compromise, because of the restrictions set by the length of the experiment. In Study 1, only five contents were examined with one type of manipulation. This restricted set was selected to show different types of contents in general as well as different aspects critical for sharpness manipulations. The results showed that IBQ-method can be used even with a manipulation that is considered artefactual and it can reveal impressions on which people base their decisions. However, the specific impressions we reported should be generalized only to a very similar material. In Study 2 examining multivariate changes due to image processing, the selection of material can be seen in the way the quality dimensions were formed: one processing produced images that were noticeably dark and the darkness dimension was the second. This choice might have emphasized the importance of

darkness dimension, but we did not consider this a problem, since illumination is one of the main image-quality characteristics. The selection of the material had to be made also in the eye tracking studies of this thesis (Studies 3 and 4), where we wanted to compare how different instructions or differences in understanding the task influenced viewing. Besides image contents also manipulations influence viewing. Therefore to reach knowledge of viewing behaviour at a general level, several versions of the same content had to be repeated. This resulted in a quite limited number of image contents. However, by selecting the image contents carefully to show different aspects common in home photography (recommended for example in I3A, 2007; Keelan, 2002), we tried to ensure the generalizability of the results. This selection was done taking into account, for example, the distance between the camera and the photographed object, whether there are people in the image, and the lighting condition in which the photograph was taken. In addition, we selected different types of processing. Like this we tried to cover many different aspects of quality estimation and to ensure that for example certain type of content or processing does not get too much power in the results. Future studies should be conducted with different types of material, especially to fully understand the estimations of different viewing groups and the result why the group with medium length fixation duration used the most abstract attributes. This examination could involve temporal aspects of viewing behaviour, since these have been shown to be an important aspect when predicting the tasks from viewing behaviour (Haji-Abolhassani & Clark, 2014; Kanan, Ray, Bseiso, Hsiao, & Cottrell, 2014; Radun, Nuutinen, Antons, & Arndt, 2016).

This thesis was restricted to quality and preference estimations related to everyday photography, and investigations in the field of art appreciation and aesthetics were excluded. For a review, see Palmer (2013), for example.

5.4 Recommendations for researchers conducting studies on subjective image-quality estimation

- Think carefully about the instructions that are given to participants, because even a minor change may alter the way they understand the task and seek information.

- Quality estimation is subjective and prone to individual differences. Use enough participants in each study.
- Examine the data for possible subgroups and try to understand why they are different.
- If possible, ask for the reasons behind the evaluations.
- With high-quality material in particular, think about the wider context the participants might have in mind when making their estimations.

5.5 Conclusions

Quality estimation in the case of high-quality images tends to be a preference task that is context-dependent and subjective, especially when the observers are naïve in this respect. The context dependency is evident in the interaction between content and quality changes, as well as more widely in participant behaviour. Factors influencing behaviour include the instructions or the participants' own conceptions of estimation rules and their expectations of the task requirements in that specific situation. The context dependency highlights the need to examine individual differences, especially when there are multivariate changes in the material. We introduced an Interpretation-Based Quality (IBQ) method that is suitable for deeper examination of participants' conceptions than standard methods. It focuses on the estimation rules on which they base their estimations in adding qualitative examination to traditional psychophysical methods of subjective image-quality estimation. Even naïve participants were able consistently to give the grounds on which they based their estimations when they could use their own language. These grounds shed light on the rules people use in the process of quality estimation.

The choice of estimation rules depends on the task content, the quality changes, the instructions and personal preferences, which also influence viewing behaviour. Attention allocation changes according to the instructions. Finally, the subjectivity of quality estimation was seen in the participants' viewing behaviour and estimation rules.

In conclusion, it has been shown in this thesis that when the quality level is high, general quality estimation is not enough to fully explain the quality process.

It is important to understand the reasons for the estimations, which with high levels of quality relate not only to low-level image features, but also to the interaction between expectations and changes expressed as differences in the meanings the image conveys.

6 References

Alers, H., Redi, J., Liu, H., & Heynderickx, I. (2015). Effects of task and image properties on visual-attention deployment in image-quality assessment. *Journal of Electronic Imaging*, *24*, 23030.

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research*, *39*, 2947–2953.

Antes, J. R. (1974). Time Course of Picture Viewing. *Journal of experimental psychology*, *103*, 62–70.

Arndt, S., Radun, J., Antons, J.-N., & Möller, S. (2014). Using Eye-Tracking and Correlates of Brain Activity to Predict Quality Scores. *QoMex 2014 The International Workshop on Quality of Multimedia Experience* (pp. 281–285). Singapore: IEEE.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature reviews. Neuroscience*, *4*, 829–839.

Bakeman, R., & Gottman, J. M. (1986). *Observing interaction: An introduction to sequential analysis*. Cambrige: Cambridge University Press.

Bartleson, C. J. (1960). Memory Colors of Familiar Objects. *Journal of the Optical Society of America*, *50*, 73–77.

Bech, S., Hamberg, R., Nijenhuis, M., Teunissen, K., Looren de Jong, H., Houben, P., & Pramanik, S. K. (1996). The RaPID Perceptual Image Description Method (RaPID). In B. E. Rogowitz & J. P. Allebach (Eds.), *Electronic Imaging: Science & Technology* (pp. 317–328). International Society for Optics and Photonics.

Beke, L., Kutas, G., Kwak, Y., Sung, G. Y., Park, D.-S., & Bodrogi, P. (2008). Color preference of aged observers compared to young observers. *Color Research & Application*, *33*, 381–394.

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research*, *25*, 187–217.

Bianco, S., Bruna, A. R., Naccari, F., & Schettini, R. (2013). Color correction pipeline optimization for digital cameras. *Journal of Electronic Imaging*, *22*, 23014.

Bleckley, M. K., Durso, F. T., Crutchfield, J. M., Engle, R. W., & Khanna, M. M. (2003). Individual differences in working memory capacity predict visual attention allocation. *Psychonomic Bulletin & Review*, *10*, 884–889.

Boot, W. R., Becic, E., & Kramer, A. F. (2009). Stable individual differences in search strategy? The effect of task demands and motivational factors on scanning strategy in visual search. *Journal of vision*, *9*, 1–16.

Boring, E. G. (1957). *A history of experimental psychology* (2nd ed.). New York, USA: Appleton-Century-Crofts.

Le Callet, P., & Niebur, E. (2013). Visual attention and applications in multimedia technologies. *Proceedings of the IEEE*, *101*, 2058–2067.

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention Perception & Psychophysics*, *72*, 1283–1297.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology-Human Perception and Performance*, *33*, 753–763.

Castelhano, M. S., & Henderson, J. M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian journal of experimental psychology = Revue canadienne de psychologie expérimentale*, *62*, 1–14.

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*, 1–15.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*, 1–15.

Chandler, D. M. (2013). Seven Challenges in Image Quality Assessment: Past, Present, and Future Research. *ISRN Signal Processing*, 1–53.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Civille, G. V., & Oftedal, K. N. (2012). Sensory evaluation techniques - Make 'good for you' taste 'good'. *Physiology and Behavior*, *107*, 598–605.

DiStefano, C., & Mindrila, D. (2013). Cluster analysis. In T. Teo (Ed.), *Handbook of Quantitative Methods for Educational Research* (pp. 103–122). SensePublishers.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*, 1–19.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*, 1–26.

Engeldrum, P. G. (2004a). A Short Image Quality Model Taxonomy. *Journal of Imaging Science and Technology*, *48*, 160–165.

Engeldrum, P. G. (2004b). A Theory of Image Quality: The Image Quality Circle. *Journal of Imaging Science and Technology*, *48*, 446–456.

Engelke, U., Kaprykowsky, H., Zepernick, H.-J., & Ndjiki-Nya, P. (2011). Visual Attention in Quality Assessment. *IEEE Signal Processing Magazine*, *28*, 50–59.

Engelke, U., Liu, H., Wang, J., Le Callet, P., Heynderickx, I., Zepernick, H.-J., & Maeder, A. (2013). Comparative study of fixation density maps. *IEEE Transactions on Image Processing*, *22*, 1121–33.

Farnand, S. (2013). *Designing pictorial stimuli for perceptual image difference experiments*. *Doctoral Dissertation*. B.S. Cornell University.

Gescheider, G. A. (1985). *Psychophysics: Method, theory and application*. Hillsdale, N.J. USA: Erlbaum Associates.

Gill, J. (2001). *Generalized Linear Models: A Unified Approach, Issue 134*. Thousand Oaks, CA:

SAGE Publications.

Greenacre, M. (2007). *Correspondence Analysis in Practice, Second Edition*. Boca Raton, USA: CRC Press.

Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting observers' task from eye movement patterns. *Vision research*, *103C*, 127–142.

Hanley, J. A. (2003). Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation. *American Journal of Epidemiology*, *157*, 364–375.

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*, 219–222.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic bulletin & review*, *16*, 850–856.

Henderson, J. M., Nuthmann, A., & Luke, S. G. (2013). Eye movement control during scene viewing: immediate effects of scene luminance on fixation durations. *Journal of experimental psychology. Human perception and performance*, *39*, 318–22.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Weijer, J. van de. (2011). *Eye Tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

I3A. (2007). *CPIQ Initiative Phase 1 White Paper: Fundamentals and review of considered test methods*.

IEC. (1999). *IEC 61966-2-1 Multimedia systems and equipment - Colour measurement and management Part 2-1: Colour management - Default RGB colour space - sRGB*.

ISO 12640-1. (1997). *ISO 12640-1 Graphic technology - Prepress digital data exchange - CMYK standard colour image data*. Geneva, Switzerland: International Organization for Standardization.

ISO 12640-2. (2004). *ISO 12640-2 Graphic technology — Prepress digital data exchange — Part 2: XYZ/sRGB encoded standard colour image data (XYZ/SCID)*. Geneva, Switzerland.

ISO 20462-1. (2005). *ISO 20462-1:2005 - Photography -- Psychophysical experimental methods for estimating image quality -- Part 1: Overview of psychophysical elements*. Geneva, Switzerland.

ISO 20462-2. (2005). *ISO 20462-2:2005 - Photography -- Psychophysical experimental methods for estimating image quality -- Part 2: Triplet comparison method*. Geneva, Switzerland.

ISO 20462-3. (2012). *ISO 20462-3:2012 - Photography -- Psychophysical experimental methods for estimating image quality -- Part 3: Quality ruler method. International standard*. Geneva, Switzerland.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, *40*, 1489–1506.

ITU-R BT.500-13. (2012). *Recommendation ITU-R BT . 500-13 Methodology for the subjective*

assessment of the quality of television pictures. Geneva, Switzerland: International Telecommunication Union.

ITU-T P.800.1. (2006). *Recommendation ITU-T P.800.1 Mean Opinion Score (MOS) terminology*. Geneva, Switzerland.

Janssen, T. J. W. M., & Blommaert, F. J. J. (2000). A computational approach to image quality. *Displays*, *21*, 129–142.

Jayaraman, D., Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). Objective quality assessment of multiply distorted images. *Asilomar Conference on Signals, Systems and Computers* (pp. 1693–1697). Pacific Grove, CA, USA: IEEE.

Judd, T., Durand, F., & Torralba, A. (2011). Fixations on low-resolution images. *Journal of Vision*, *11*, 1–20.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to Predict Where Humans Look. *IEEE International Conference on Computer Vision*, 2106–2113.

Kaller, C. P., Rahm, B., Bolkenius, K., & Unterrainer, J. M. (2009). Eye movements and visuospatial problem solving: identifying separable phases of complex cognition. *Psychophysiology*, *46*, 818–30.

Kanan, C., Ray, N. A., Bseiso, D. N. F., Hsiao, J. H., & Cottrell, G. W. (2014). Predicting an observer's task using multi-fixation pattern analysis. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '14* (pp. 287–290). New York, New York, USA: ACM Press.

Kao, W., Wang, S., Chen, L., & Lin, S. (2006). Design considerations of color image processing pipeline for digital cameras. *IEEE Transactions on Consumer Electronics*, *52*, 1144–1152.

Keelan, B. W. (2002). *Handbook of Image Quality - Characterization and Prediction*. New York, USA: Marcel Dekker, Inc.

Keelan, B. W., & Urabe, H. (2004). ISO 20462: a psychophysical image quality measurement standard. In Y. Miyake & R. Rasmussen (Eds.), *Proc. SPIE 5294: Image Quality and System Performance* (Vol. 5294, pp. 181–189). San Jose, CA, USA: International Society for Optics and Photonics.

Knudsen, E. I. (2007). Fundamental components of attention. *Annual review of neuroscience*, *30*, 57–78.

Kortum, P., & Geisler, W. S. (1996). Implementation of a foveated image coding system for image bandwidth reduction. In B. E. Rogowitz & J. P. Allebach (Eds.), *Proc. SPIE 2657, Human Vision and Electronic Imaging* (pp. 350–360). San Jose, CA, USA: International Society for Optics and Photonics.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological review*, *118*, 97–109.

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Progress in retinal and eye research*, *25*, 296–324.

Land, M. F. (2009). Vision, eye movements, and natural behavior. *Visual neuroscience*, *26*, 51–

62.

Larson, E. C., Vu, C., & Chandler, D. M. (2008). Can visual fixation patterns improve image fidelity assessment? *2008 15th IEEE International Conference on Image Processing (ICIP 2008)* (pp. 2572–2575). San Diego, CA, USA: IEEE.

Lindemann, L., & Magnor, M. (2011). Assessing the quality of compressed images using EEG. *2011 18th IEEE International Conference on Image Processing (ICIP 2011)* (pp. 3109–3112). Brussels, Belgium: IEEE.

Liu, H., Engelke, U., Le Callet, P., & Heynderickx, I. (2013). How Does Image Content Affect the Added Value of Visual Attention in Objective Image Quality Assessment? *IEEE Signal Processing Letters*, *20*, 355–358.

Liu, H., & Heynderickx, I. (2011). Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data. *IEEE Transactions on Circuits and Systems for Video Technology*, *21*, 971–982.

Meilgaard, M. C., Civille, G. V., & Carr, B. T. (1999). *Sensory Evaluation Techniques* (3th ed.). Boca Raton, FL, USA: CRC Press.

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision*, *11*, 1–15.

Ninassi, A., Le Meur, O., Le Callet, P., Barba, D., & Tirel, A. (2006). Task impact on the visual attention in subjective image quality assessment. *14th European Signal Processing Conference (EUSIPCO 2006)*. Florence, Italy: EURASIP.

Nuutinen, M., Virtanen, T., Leisti, T., Mustonen, T., Radun, J., & Häkkinen, J. (2016). A new method for evaluating the subjective image quality of photographs: dynamic reference. *Multimedia Tools and Applications*, *75*, 2367–2391.

Nyman, G., Radun, J. E., Leisti, T., & Vuori, T. (2005). From image fidelity to subjective quality: a hybrid qualitative/quantitative methodology for measuring subjective image quality for different image contents. *Proceedings of the 12th International Display Workshops (IDW'05)* (pp. 1817–1820). Takamatsu, Japan.

Nyman, G., Radun, J., Leisti, T., Oja, J., Ojanen, H., Olives, J.-L., Vuori, T., et al. (2006). What do users really perceive - probing the subjective image quality. In L. Cui & Y. Miyake (Eds.), *Volume 6059 Image Quality and System Performance III* (Vol. 6059, pp. 605902-1–7). San Jose, CA, USA: International Society for Optics and Photonics.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*, 520–527.

Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia*, *45*, 75–92.

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual review of psychology*, *64*, 77–107.

Payne, J. W., Bettman, J. R., & Schkade, D. A. (1999). Measuring constructed preferences: Towards a building code. *Journal of Risk & Uncertainty*, *19*, 243–270.

Pedersen, M., Bonnier, N., Hardeberg, J. Y., & Albregtsen, F. (2010). Attributes of image quality for color prints. *Journal of Electronic Imaging*, *19*, 11016.

Radun, J., Nuutinen, M., Antons, J.-N., & Arndt, S. (2016). Did you notice it? – How can we predict the subjective detection of video quality changes from eye movements? *IEEE Journal of Selected Topics in Signal Processing*, 1–11.

Radun, J., Virtanen, T., & Nyman, G. (2006). Explaining multivariate image quality - Interpretation-Based Quality Approach. *ICIS '06: International Congress of Imaging Science* (pp. 119–121). Rochester, NY, US: IS&T.

Ramanath, R., Snyder, W. E., Yoo, Y., & Drew, M. S. (2005). Color image processing pipeline. *IEEE Signal Processing Magazine*, *22*, 34–43.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*, 1457–1506.

Rayner, K., Li, X., Williams, C. C., Cave, K. R., & Well, A. D. (2007). Eye movements during information processing tasks: individual differences and cultural effects. *Vision research*, *47*, 2714–26.

Salmi, H., Halonen, R., Leisti, T., Oittinen, P., & Saarelma, H. (2009). Development of a balanced test image for visual print quality evaluation. In S. Farnand & F. Gaykema (Eds.), *SPIE Proceedings Vol. 7242: Image Quality and System Performance VI* (Vol. 7242, pp. 7210–7242). International Society for Optics and Photonics.

Scholler, S., Bosse, S., Treder, M. S., Blankertz, B., Curio, G., Müller, K.-R., & Wiegand, T. (2012). Toward a direct measure of video quality perception using EEG. *IEEE Transactions on Image Processing*, *21*, 2619–29.

Segur, R. K. (2000). Using Photographic Space to Improve the Evaluation of Consumer Cameras. *PICS 2000: Image Processing, Image Quality, Image Capture, Systems Conference* (pp. 221–224). Portland, OR, USA: IS&T.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–726.

Strauss, A. L., & Corbin, J. M. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. *Sage* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Teunissen, K. (1996). The Validity of CCIR Quality Indicators Along a Graphical Scale. *SMPTE Motion Imaging Journal*, *105*, 144–149.

Tinio, P. P. L., Leder, H., & Strasser, M. (2011). Image quality and the aesthetic judgment of photographs: Contrast, sharpness, and grain teased apart and put together. *Psychology of Aesthetics, Creativity, and the Arts*, *5*, 165–176.

To, M. P. S., Lovell, P. G., Troscianko, T., & Tolhurst, D. J. (2010). Perception of suprathreshold naturalistic changes in colored natural images. *Journal of Vision*, *10*, 1–22.

Tolhurst, D. J. (2013). *Workshop presentation*. University of Helsinki.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America. A, Optics, image science, and vision*, *20*, 1407–1418.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*, 97–136.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*, 1395–1407.

Wang, Z., & Bovik, A. C. (2001). Embedded foveation image coding. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, *10*, 1397–410.

Warren, C., Mcgraw, A. P., & Van Boven, L. (2011). Values and preferences: Defining preference construction. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 193–205.

Virtanen, T., Nuutinen, M., Vaahteranoksa, M., Oittinen, P., & Häkkinen, J. (2015). CID2013: a database for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, *24*, 390–402.

Vu, C. T., Larson, E. C., & Chandler, D. M. (2008). Visual Fixation Patterns when Judging Image Quality: Effects of Distortion Type, Amount, and Subject Experience. *2008 IEEE Southwest Symposium on Image Analysis and Interpretation* (pp. 73–76). IEEE.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum Press.

Yendrikhovskij, S. N., Blommaert, F. J. J., & de Ridder, H. (1999). Color reproduction and the naturalness constraint. *Color Research & Application*, *24*, 52–67.

Zeng, X., Ruan, D., & Koehl, L. (2008). Intelligent sensory evaluation: Concepts, implementations, and applications. *Mathematics and Computers in Simulation*, *77*, 443–452.

Zhou, J., & Glotzbach, J. (2007). Image Pipeline Tuning for Digital Cameras. *2007 IEEE International Symposium on Consumer Electronics* (pp. 1–4). Irving, TX, USA: IEEE.