University of Twente

Master Thesis

Using mixed-effects modeling to account for the acquiescence response style bias in HCI research

Author:
Inga Schwabe

Supervisor:
Martin Schmettow

A thesis submitted in fulfilment of the requirements for the degree of Master of Sciences

in

Human Factors and Media Psychology
Psychology

January 2013

Declaration of Authorship

I, Inga SCHWABE, declare that this thesis titled, 'Using mixed-effects modeling to account for the acquiescence response style bias in HCI research' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a master degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

UNIVERSITY OF TWENTE

Abstract

Social Sciences
Psychology

Master of Sciences

Using mixed-effects modeling to account for the acquiescence response style bias in HCI research

by Inga SCHWABE

Many Human Computer Interaction (HCI) studies use questionnaires with multiple Likert scales as measurement tool. However, it is well known in the statistical literature that such measures are often biased by the influence of response styles. One of these response styles is the acquiescence response style (ARS) − defined as the *'disproportionate use of positive response options'*(Weijters, Geuen, & Schillewaert, 2010, p. 1). The impact of this response style has to be taken seriously. For example, means become non-interpretable and correlations can be found that do not reflect reality. The purpose of this thesis was twofold: 1) Show that the influence of the ARS is a threat to the validity of HCI research results and 2) Show that mixed-effects modeling can be a reliable tool to account for its impact. By replicating the study by Hassenzahl and Monk (2010), it was shown that a mixed-effects model can indeed correct for the impact of the ARS. However, contrary to our expectations the influence of the ARS was so small that it can be neglected. Furthermore, the results of an analysis on the psychometric level suggest that the scales "beauty" and "hedonic quality" are indistinguishable and therefore measure the same underlying latent variable. In the discussion section, possible explanations and directions for further research are provided.

Contents

List of Figures

List of Tables

Abbreviations

ARS	Acquiescence response style
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit index
CI	Confidence Interval
DARS	Disacquiescence response style
EFA	Exploratory factor analysis
ERS	Extreme response style
LCFA	Latent-class confirmatory factory analysis
MIRS	Mid-point response style
MRS	Middle response style
MSEP	Mean square error of prediction
MTMM	Multi-trait-multi-method models
NARS	Net acquiescnece response style
NNFI	The Nonnormed Fit index
PCA	Principal component analysis
RIRMACS	Representative indicators response styles means and covariance structure
RIRS	Representative indicators for response styles
RMSEA	Root Mean Square Error of Approximation
RS	Response style
SRMR	The Standardized Root Mean Square Residual
SS	Sum of squares
Std Dev	Standard deviation
Subj	Subjects

Symbols

$\boldsymbol{\beta}$	Population values of regression coefficients (with appropriate subscripts as needed); regression of one endogenous construct on another endogenous construct (SEM notation)
$\boldsymbol{\chi^2}$	A statistical test based on the chi-square distribution
$\boldsymbol{\sigma}$	Population standard deviation
$\boldsymbol{\sigma}^2$	Population variance
$\boldsymbol{\phi}$	Covariance (SEM notation)
$\boldsymbol{\delta}$	Measurement error associated with X measure (SEM notation)
$\boldsymbol{\epsilon}$	Measurement error associated with Y measure (SEM notation)
$\boldsymbol{\xi}$	Exogenous construct (SEM notation)
$\boldsymbol{\lambda}$	Factor loading (SEM notation)

Glossary

Akaike information criterion The Akaike information criterion (AIC) is a measure of the relative goodness of fit of a particular statistical model. It represents a relative measure of the information lost when the particular model is used to predict the outcome variable. Therefore, the AIC should be as low as possible. xvii, 29, 37, 39, 46

Collinearity Collinearity is a problem in multiple regression that can be seen when one or more (Multicollinearity) of the independent variables are highly correlated with each other. xvii, 29

Confirmatory factor analysis Confirmatory factor analysis (CFA) is used to verify the assumed factor structure of a set of observed variables. CFA allows to test the hypothesis that there is indeed a relationship between the observed variables and an underlying latent construct. Therefore, contrary to EFA, it is confirmatory and not exploratory. xvii

Discriminant validity Discriminant validity means that a test of a concept is not highly correlated with other tests designed to measures theoretically different concepts. xvii, 46

Exploratory factor analysis The exploratory factor analysis (EFA) represents a simplification of interrelated measures. Unlike PCA, it is not used to reduce data, but to explore the possible underlying factor structure of a set of variables. xvii, 36, 37

Heteroscedasticity The error terms do not have constant variance. This means, that the scatter of errors is different, varying depending on the value of one or more independent variables. Technically, this means $V\epsilon_j \neq \sigma^2$ for all j. xvii, 17

Principal component analysis Principal component analysis (PCA) is a mathematical procedure which uses a orthogonal transformation to convert observation of probable correlated variables into a set of values of linearly uncorrelated variables ("principal components"). xvii, 36, 37, 46

Structural equation modeling Structural equation modeling (SEM) is a statistical technique which purpose is to test and estimate causal relations by using a combination of statistical data and qualitative causal assumptions. xvii, 23

Variance inflation factor The variance inflation factor (VIF) represents the severity of collinearity in a regression analysis. It provides an index of the increase of the variance of an estimated regression coefficient because of collinearity. The VIF factor for $\hat{\beta}_i$ is calculated with the following formula: VIF $= \frac{1}{1-R_i^2}$. xvii, 33

Chapter 1

Introduction

1.1 Human Computer Interaction

Human Computer Interaction (HCI) is quite a new research field that is still developing. It concerns the study, planning and design of the interaction between humans (computer users) and computers. This interaction is generalized by the means of a user interface, which can be seen as a mediator between the human and the computer (Helander, 1988). HCI is often referred to as the intersection of behavioural sciences, computer science, design and several other research fields (Moran, Card, & Newell, 1983).

The ultimate goal of HCI is to improve the quality of interactions between computer users and computers by improving the usability and receptiveness of computers. Especially, HCI is concerned with the development of (Sokolowski & Banks, 2010):

- Methodologies and processes for the design of interfaces
- Methods for the implementation of interfaces
- Techniques for the evaluation and comparison of interfaces
- New interfaces and interaction techniques
- Descriptive and predictive models and theories of interaction

A long term goal of HCI is the development of systems that minimize the barrier between the human's cognitive model of what they want to accomplish and the computer's understanding of the user's task.

The research field of HCI emerged in the early 1980s, initially as an area in computer science (Carroll, 2009). In the beginning of this new discipline, there were

only a handful of experts who were devoted primarily to hardware-oriented research such as the design of input devices and CRT(Cathode ray tube) screens (Helander, 1988). HCI expanded rapidly and steadily for three decades (Carroll, 2009) and the interest shifted towards principles for information presentation. Nowadays, human formation processing and cognition supply the foundation for the main part of the research. Also, most computer manufactures and many software companies have a human factors staff with HCI experts by now.

1.1.1 Research in HCI

As the improvement of usability and receptiveness of computers is seen as the ultimate goal of HCI, much HCI research is done in the usability field. As stated by the ISO norm ISO 9241-11 (1998), usability can be defined as

- the extent to which a product
- can be used by specified users
- to achieve specified goals
- with effectiveness, efficiency and satisfaction
- in a specified context of use.

There are many different methods that can be used to measure usability. For example, one can think of controlled experiments, eye tracking, task analyses, simulation or think-aloud methods. In his paper 'Current practice in measuring usability: Challenges to usability studies and research', Hornbaek (2006) reviews current practice in measuring usability in HCI research by discussing usability measures from 18 studies published in core HCI jorunals and proceedings. He outlines several problems with the common used measures. Among others, the author emphasizes the need to study *both* objective and subjective measures. However, practice shows that many HCI researchers draw their conclusions from Likert scales only, which is a subjective measure.

1.1.2 Likert scales

The *Likert scale* is one of the most common used item formats. When a Likert scale is used, the items are presented as a declarative statement, followed by response options that indicate varying degree of agreement with or endorsement of the given statement

(de Vellis, 2003). The number of possible response options can vary, depending on the phenomenon investigated and the particular research aim. A common practice is to include five possible responses (also referred to as 5-point Likert scale): "*strongly disagree*", "*disagree*", "*neither agree nor disagree*", "*agree*", and "*strongly agree*". For the final data analysis, these responses are scored as 5 (*strongly agree*), 4 (*agree*), 3 *(neither agree nor disagree)*, 2 (*disagree*) and 1 (*strongly disagree*).

After the respondents have completed the questionnaire, each Likert scale item can be analysed individually or in some cases item responses can be summed to create a score for a group of items (de Vellis, 2003).

However, in the statistical literature, Likert scales are often criticized (e.g. Flaskerud, 1988; de Vellis, 2003; S.Jamieson, 2004; Carifio & Perla, 2007). One point of criticism is the fact that Likert scales are often biased by response styles. In this thesis, we focus on the impact of this bias. The reason for focusing on response styles is that even though these errors are discussed elaborately in the methodological literature (e.g. Dooley, 2001; de Vellis, 2003; Kaplan & Saccuzzo, 2009) and can have important ramifications for the conclusions derived from HCI studies (elaborated in Section 1.2.4), HCI researchers seldom do pay sufficient attention to them.

1.1.3 Semantic differential rating scale

There is another scale type that is commonly used in usability testing: The semantic differential rating scale (e.g. in Hassenzahl & Monk, 2010; Tractinsky, 1997; Kurosu & Kashimura, 1995). The scale was invented by Osgood and designed to measure the connotative meaning of concepts (Osgood & Tannenbaum, 1957). In this item format, the respondent has to choose where his position lies on a scale between two bipolar adjectives (e.g. "Good-Bad" or "Happy-Unhappy"). In usability research, this sort scale is often used to measure the usability of a website. You can see an example of such a scale in Figure 1.1.

In the example shown in Figure 1.1, the participant has to evaluate the usability of a website by the means of seven different semantic differential rating scales. He is asked to indicate his position for each dimension by putting a mark on one of the seven spaces along each dimension.

For the analysis of a semantic differential rating scale, the scoring of the data follows the same principle as the scoring for a Likert scale item (e.g. when the first space

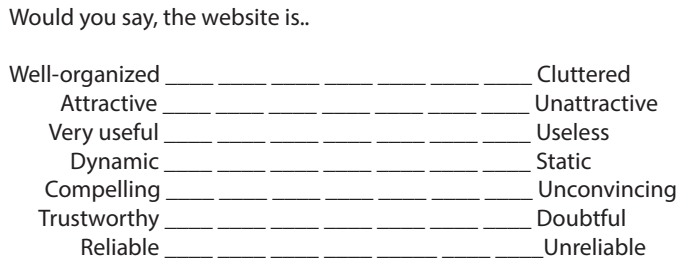

FIGURE 1.1: Example of a semantic differential rating scale

is marked, a 1 is scored and when the fifth space is marked, a 3 is scored). Therefore, the methodological problem of a Likert scale, the possible influence of a response style bias, can also be found in studies that use this kind of scale as well. This was another motivation to focus on the impact of response styles. In the following section, a definition of response styles, their sources and their impact is provided.

1.2 Response styles

A response style (RS) is defined as respondent's behavioural tendencies to disproportionately select a subset of the available response options (Rorer, 1965; Weijters, Geuen, & Schillewaert, 2010). It refers to a person's manner of responding to test items, independent of the item content. There are some researchers who view a response style as a personality trait. This would mean that it appears on different tests regardless of their content and that it persists over test occasion (Dooley, 2001).

1.2.1 Types

Two response styles in particular have received the focus of attention in behavioural sciences research: The acquiescence response style bias (ARS) and the extreme response style (ERS) (Johnson, Kulsea, Cho, & Shavitt, 2005; van Herk, Poortinga, & Verhallen, 2004).

ERS is the tendency of a respondent to choose the most extreme categories of the given rating scale. Practically, this means that a high-ERS participant who is given a 7-point Likert type scale will tend to respond with 1 (*strongly disagree*) or 7 (*strongly disagree*). On the other hand, if a low-ERS participant is given the same survey, his or her responses will tend to cluster around 4 (*neither disagree or agree*)(Cheung & Rensvold, 2000). There are several studies that have documented cross-cultural differences in ERS

(e.g. Hui & Triandis, 1985; Schaninger & Buss, 1986; Greenleaf, 1992; Triandis, 1994). Research by Lee and Green (1991) for example suggests that Koreans tend to avoid extremes and prefer the midpoints of scales.

ARS refers to the *'disproportionate use of positive response options'* (Weijters, Geuen, & Schillewaert, 2010, p. 1). Practically, this means that the response of a high-ARS respondent will show a pattern of reflexively agreeing with survey items. If a low-ARS participant is given the same survey, he will show a cluster of sometimes agreement and sometimes disagreement with the statements of the survey. There are several studies that have documented cross-cultural differences in ARS (e.g. Cunningham, Cunningham, & Green, 1977; England & Harpaz, 1983; Morris & Pavett, 1992; Riordan & Vandenberg, 1994). Research by Riordan and Vandenberg (1994) for example shows that a response of 3 on a 5-point Likert-type sclae means "no opinion" to American Respondents but "mild agreement" to Korean respondents. As a result of these different interpretations, Korean "3"s were equivalent to American "4"s and Korean "4"s were equivalent to American "5"s. Differences in ARS can be explained in terms of *social desirability* - the belief that a higher score is a better score (Guilford, 1954; Hui & Triandis, 1985; Moorman & Podaskoff, 1992; Peterson & Wilson, 1992).On the individual level, there can be some respondents who display extreme ARS by agreeing (or disagreeing) with almost any statement (Guilford, 1954; Peterson & Wilson, 1992; Triandis, 1994).

Both ERS and ARS differences can be either *nonuniform* or *uniform* (Cheung & Rensvold, 2000).

1.2.2 Sources

Sources of response styles can be classified into two main categories: the stimulus level and the respondent level. At the stimulus level, response styles are seen as a consequence of the survey. At the respondent level, response styles are seen as a consequence of personal characteristics of the respondent (B.Weijters, 2006).

In the current study, we will concentrate on the respondent level. Factors that can play a role on the respondent level are demographic variables (education, age, gender, income and employment), personality variables and culture- and country-level characteristics. Overall, demographic and personality variables explain a quite small proportion of the variance of response styles, whereas culture and country-level characteristics are found to explain a relatively large proportion of response styles (van Vaerenbergh &

Thomas, 2012). For example, Weijters, Geuens, and Schillewaert (2010) found in a study using a Belgian sample that demographic variables explain between 1.4 % and 8.3% of the variance in response styles depending on which response style is considered, whereas Meisenberg and Williams (2008) found that socio-demographic variables (e.g. corruption or gross domestic product) explain only 1-5 % of the variance in ARS and ERS at the individual level but country characteristics explain 63.2 % (ARS) to 74.5 % (ERS) at the country level.

1.2.3 Methods of detecting and correcting for response styles

The literature identifies several ways to detect and control response styles. In a literature review, van Vaerenbergh and Thomas (2012) compared these different techniques. For an overview of the different approaches see Table 1.1.

Methods of detecting and correcting for response styles

Measurement of RS	Description	Advantages	Disadvantages	Representative studies
Count procedure	Across an entire questionnaire, count the number of agreements, disagreements, extreme responses, and/or mid-point responses	Easy to use, no additional indicators needed	The method only works with heterogeneous items	(Bachman & O'Malley, 1984; N. Reynolds, 2010)
Counting double agreements on reversed item	Reversed items are included in the questionnaire, and later, the number of double agreements on the reversed items is counted	Easy to use, additional indicators are not necessary	It is sometimes difficult to interpret people's responses to reversed items, they might e.g. also due to interpretational factors	(Hox, Leeuw, & Kreft, 1991; Johnson, Kulesa, Cho, & Shavitt, 2005)
MTMM	Repeated measurement of the same trait using different methods. The observed variance can then be split up into true variance and error variance	The method is easy to set up and use. It measures net effects of ARS and DARS and there are no additional indicators needed	Method gives no indication of ERS and MRS. Due to repeated measurement, consistency bias and memory effect might arise. Last, problems of identification arise often	(Saris & Aalberts, 2003; Saris, Satorra, & Coenders, 2004)

Methods of detecting and correcting for response styles

Measurement of RS	Description	Advantages	Disadvantages	Representative studies
Specify method factor in CFA	Positive & negative loadings on a content factor and positive loadings on a method factor are specified	It is relatively easy to specify (most researchers are familiar with CFA) and no additional indicators are needed	The method does not control for DARS, MRS, or ERS and requires the use of balanced scale items. All loadings on the method factor have to be restricted to equality in order to identify the model	(Billiet & McClendon, 2000; Welkenhuysen-Gybels, Billiet, & Cambr, 2003)
Latent-class regression analysis	A latent-class regression analysis has to be runned and one has to assess whether a method factor emerges	No additional indicators are needed	Specific software has to be used. Researchers might be unfamiliar with latent-class analysis. Sometimes hard to specify	(Moors, 2010; van Rosmalen, van Herk, & Groenen, 2010)
LCFA	Two method factors are specified, one to measure ARS and one to measure ERS	No additional indicators are neded. Recent models allow discriminating ARS and ERS	The method does not account for DARS and MRS and specific software has to be used. Researchers might be unfamiliar with LCFA	(G.Moors, 2003, 2012; Kieruj & Moors, 2012)

Methods of detecting and correcting for response styles

Measurement of RS	Description	Advantages	Disadvantages	Representative studies
IRT Model	The method models the probability of choosing a particular response option as a function of the underlying latent variable	It allows different items to be differentially useful for measuring ERS. Relaxes the assumption that ERS measures have to be uncorrelated	Method is only developed for ERS. It requires use of Markov chain Monte Carlo procedures , which might be difficult to implement for some researchers	(Bolt & Newton, 2011; de Jong, Steenkamp, Fox, & Baumgartner, 2008)
RIRS method	A number of uncorrelated, maximally heterogeneous measures in content to the survey are included and the weighted RS indicators are calculated	Easy to calculate. Allows the measurement of ARS, DARS, ERS, MRS, NARS. It is not related to content and easy to include as covariates in subsequent analyses	The researcher has to add additional items to the survey	(Baumgartner & Steenkamp, 2001; Greenleaf, 1992; B.Weijters, 2006)

Methods of detecting and correcting for response styles

Measurement of RS	Description	Advantages	Disadvantages	Representative studies
RIRMACS method	Additional, uncorrelated items are added to the survey, which serve as observed variables in a CFA, ARS, DARS, MRS and ERS as latent variables. It extend the RIRS method	Easy to use. RS indicators can be added as covariates in subsequent analyses. Makes use of specific RS indicators which allows discrimination between content and style. Allows measurement of ARS, DARS, MRS and ERS. Allows testing of convergent and discriminant validity of the different RS	The researcher has to add additional items to the survey	(Weijters, Geuens, & Schillewaert, 2008)

Source: (van Vaerenbergh & Thomas, 2012). For a list of the abbreviations, see 1

However, the author also points out some difficulties of some of these methods.

First of all, the methods *double agreements n reversed items* (Johnson, Kulesa, et al., 2005) and *specifying a method factor on balanced-scale items* (Billiet & McClendon, 2000) require the use of balanced-scale items. This can be problematic, because it is often difficult to formulate reversed items (Billiet & McClendon, 2000). Furthermore, the way people respond to reversed items may be due to interpretational issues and not reflect a response style bias (Wong, Rindfleisch, & Burroughs, 2003). For example, Weijters, Geuens, and Schillewaert (2009) state that respondents tend to minimize retrieval of additional information when they have to answer nearly non reversed items, but tend to maximize retrieval of new/different information when they have to answer nearby reversed items. As a result, balanced scales cause several other threats to the validity of the research results. Moreover, these methods may not always be applicable, because the majority of the most common scales are not balanced (Baumgartner & Steenkamp, 2001).

Second, not all methods account for multiple types of response styles. For example, the Multi-trait-multi-method model accounts for the ARS and the disacquiescence response style (DARS) but not for the ERS or the middle response style (MRS) (Saris & Aalberts, 2003). The method that can cover the most response styles is to add representative indicators for response styles (RIRS) to the questionnaire, which allows the calculation of ARS, DARS, ERS, MRS and the net acquiescence response style (NARS) (Baumgartner & Steenkamp, 2001; Weijters et al., 2008). In regular studies it is recommended to include five items per response style and in studies explicitly focusing on response styles, 10-14 items should be included (Weijters et al., 2008). This is probably not always possible because of survey length restrictions (van Vaerenbergh & Thomas, 2012).

Third, the convergent validity between the different methods is not well established. In their paper Beuckelaer, Weijters, and Rutten (2010) compare the RIRS method with the more traditional method (count procedure) in which survey items are also used to model ARS and ERS. Although the proportion of ARS is the same for both methods, the correlation between the methods is low to very low. In contrast, the proportion of ERs is higher for the traditional method (count procedure), but the correlation between both methods is moderate to strong.

The authors van Vaerenbergh and Thomas (2012) recommend different methods for different situations and finish their review with the conclusion that researchers should

use multiple methods to account for response styles and to assess the stability of their findings across the methods.

1.2.4 Impact

The impact of these response styles has to be taken seriously. They have huge impact on a research's reliability and validity.

First of all, estimates of the means of observed variables in a given sample can be misleading (Baumgartner & Steenkamp, 2005). Furthermore, as the response styles affect numerical scores, comparisons of means become non interpretable (Cheung & Rensvold, 2000). Consider the following study as an example for the illustration of this problem. Baumgartner and Steenkamp (2001) compared in their study consumer's attitude toward advertising in Denmark, France, Netherlands, and Portugal. Here for, they had large, nationally representative samples of consumers from all countries complete a five-item attitude toward advertising scale. The participants had to rate these five items on a 5-point Likert scale. When analysing the unadjusted means of these five observed variables, they found that the differences were small and non significant. However, when accounting for the bias, the means were significantly different from each other at $p < 0.001$. Portuguese consumers had significantly more negative attitudes than Denmark and France, and Dutch consumers held the most negative attitudes of all.

Second, estimates of the relationship between observed variables can be misleading. Consider the following study as an example of this problem. Baumgartner and Steenkamp (2001) studied the relationships between health consciousness (HCO), quality consciousness (QCO), environmental consciousness (ECO), and consumer ethnocentrism (CET) across more than 10.000 consumers in 11 countries of the European Union. When correlating the observed measures for these four variables, the pairwise correlations were all significant and sometimes substantial: HCO-QCO 0.40; HCO-ECO 0.33; HCO-CET 0.28; QCO-ECO 0.31; QCO-CET 0.19; and ECO-CET 0.15. However, when they conducted an extensive analysis of response styles, they found that all four scales were contaminated by construct-irrelevant response style variance. The mean percentage of total variance was accounted for by five different response styles. When they removed variance due to stylistic responding, the correlations between the scales were

greatly reduced: HCO-QCO 0.20; HCO-ECO 0.15; HCO-CET 0.02; QCO-ECO 0.13; QCO-CET <0.001; ECO-CET 0.01.

Third, on a more fundamental level, response styles influence factor loadings and intercepts. Therefore, the numbers on the response scale mean different things to members of different groups. Whereas ERS affects both factor loadings and intercepts, ARS affects only intercepts Cheung and Rensvold (2000).

1.2.5 Impact in HCI research

As shown in Section 1.2.4, research suggests that response styles have severe impact on the results of marketing and consumer research. In the HCI research field, there has not been much research about response styles (in a literature study, not one single article could be found). Therefore, knowledge is limited and only assumptions can be made about the possible impact of response styles in HCI research. However, review of methodology and research results of some typical HCI research disciplines show that these are quite vulnerable for one response style in particular: The ARS. In the following, we will discuss the possible influence of the ARS in two HCI research disciplines.

One of these research disciplines is research that makes use of the Technology Acceptance Model (TAM). TAM is a model very well-known in information systems research. The model predicts how users come to accept and use a technology (Davis, 1986). The model suggests that in particular two factors influence the user's decision about how and when he will use a new technology:

- Perceived usefulness (PU) - Defined as the degree to which a user believes that the use of a particular system is enhancing his job performance
- Perceived ease-of-use (PEIOU) - Defined as the degree to which a user believes that using a particular system is effort free

To measure these two variables, Davis (1993) constructed two Likert scales of which each contains ten items.

The study by Hsi-Peng, Huei-Ju, and Simon (2001) can be named here as an example of a research that makes use of this model. In their study, they present a path analytic model of people's willingness to use different computerized models from the perspectives of individual's cognitive styles, beliefs and attitudes. Among other variables, they also measured the two constructs of the TAM. Furthermore, among other

research goals, they were interested in the relationship between perceived usefulness and people' willingness to use the systems and the relationship between perceived ease-of-use and people' willingness to use the systems.

They found that, consistent with previous studies, perceived usefulness had a strong influence on people' willingness to use the systems (r=0.65), while perceived ease-of-use had a smaller but still significant effect (r=0.33), but they did not account for the possibility that these found correlations might be influenced by a response style bias. The aforementioned studies (see Section 1.2.4) suggests that it is likely that participant's answers are influenced by the ARS. Remember that ARS refers to the *'disproportionate use of positive response options'*(Weijters, Geuen, & Schillewaert, 2010, p. 1), which practically means that the response of a high-ARS respondent will show a pattern of reflexively agreeing with survey items. Practically, this means that when the respondents of these researchers show an ARS pattern, they tend to choose the right end of the Likert-scale items (*agree* or *totally agree*). As response styles are quite constant over time (see Section 1.2), it is likely that the respondents who show an ARS pattern would do this on *every* Likert scale item they are asked to answer in the particular research. Therefore, in this research discipline, respondents would show an ARS pattern not only for the scale that measures for example perceived usefulness but also for the other scales. Therefore, the found correlations could be artefact of the ARS in both scales, as not only the true score of the respondent is measured, but also the ARS. When this measurement would be corrected for the ARS, the correlation could be less strong or even totally fade away.

Another research discipline that is particularly threatened by the influence of ARS is research about the relationship of beauty and usability. As this particular HCI research field was the focus of the current study, we will take a closer look at some studies and investigate the methodology and research design in order to explore their potential to account for a possible impact of the ARS.

1.2.5.1 Influence of ARS in beauty and usability studies

One HCI research field focuses on one particular aspect of usability (as defined by the ISO norm ISO 9241-11 (1998), see Section 1.1.1 to review it), that is, the satisfaction of the user. In the HCI research literature, this research field is generally referred to as user experience (UX) research. It is an approach to HCI which "emphasizes subjectively experienced, positive, and non instrumental outcomes of owning and using

interactive products as complement to the traditional, predominantly task-oriented approach" (Hassenzahl & Monk, 2010, p.2). Over the last years, the study of aesthetics, that is, beauty, has become part of the UX research field. One particular research aim consists of investigating the relationship between perceived beauty and perceived usability.

Most studies in the beauty and usability research field are correlative (Tucha, Rotha, Hornbaek, Opwisa, & Bargas-Avilaa, 2012). That means that the variables are not systematically manipulated as independent experimental factors. The typical procedure of a study of this HCI research field is that participants have to rate the perceived usability and the perceived beauty of a product. As a measurement, two Likert scales are used, one for the perceived usability and one for the perceived beauty of the participant. Then these two measures are correlated. From that point on, these studies run all into the same problem.

Many of these studies (e.g. Chawda, Craft, Cairns, Rger, & Heesch, 2005; Porat & Tractinsky, 2012) found a positive relationship between these two constructs, but not one did account for the possibility that a response style bias influences the research results. For the aforementioned reason (see Section 1.2.5), the high correlations that were found could be an artefact of the ARs rather than reflect a true relationship. The methodology and research design of these studies allow the researcher to account for the ARS to a different degree respectively. In the following, each study and its potential to account for the ARS will be discussed separately.

In the first study, Chawda et al. (2005) tested the relationship between user perceptions of aesthetics and usability to evaluate Norman's assertion that "attractive things work better" (Norman, 2002). Participants had to rate aesthetics and usability on a Likert scale questionnaire prior and after each test run with a record kept of performance. Pre-use and post-use measures indicated that there was a strong relationship between participant's judgements of aesthetics and usability. There was no association found between the performance of participants and their rated aesthetics. The correlation between the two Likert scales lead the authors to the conclusion that attractive things are perceived to work better. In this study, participants had to rate the aesthetics and usability of one search tool visualisation only. Therefore, it is not possible to estimate the specific ARS for each person (see Section 1.3.3 for an elaborated explanation). This leads to the conclusion that the research design of this study makes it impossible to account for the influence of the ARS.

In the second study, Porat and Tractinsky (2012) developed and tested a model that suggests that salient design characteristics (aesthetics and usability) of a web store influence the emotions of the store's site visitors, which in turn affect their attitude toward the store. They found that the effect of the design aspects on attitude towards the store was partially mediated by affect. Additionally to this, they found that certain design aspects affected attitudes directly − shown by their high inter item correlations. Again, there were no repeated measures taken; participants had to evaluate one web store only. Therefore, it is be possible to account for the influence of the ARS in this research design.

The last study by Hassenzahl and Monk (2010) focused on the interplay between usability, goodness and beauty. In their study, participants had to evaluate different websites. Therefore, one can account for the possible effect of the ARS (see Section 1.3.3 for an elaborated explanation). As the current study is a replication of this study, its research design will be explained elaborately in Section 1.4.

Because of the- special vulnerability of HCI studies for the influence of the ARS, the current study focused on this particular response style. The lack of studies researching this problem shows that it is an important research goal to discover whether it is indeed a problem and if so, to find methods to account for this problem.

1.3 Alternative method for correcting for ARS

In this thesis, we introduce a new method to account for the influence of ARS: Mixed effects modeling. To explain the link between response styles and mixed-effects modeling we will first provide a small introduction to mixed-effects modeling. To give the reader a better understanding, first of all the concepts of fixed and random effects will be explained shortly.

1.3.1 Fixed and random effects

The terms 'random effects' and 'fixed effects' are used in the context of ANOVA and regression models. They refer to a certain type of statistical model.

A fixed effect is a statistical model which assumes that an independent variable is fixed. Such a model is used when it is assumed that the generalization of the results apply to similar values of independent variables in the population or in other studies. Fixed effects can be thought of as "treatment" levels that a researcher has selected for

inclusion in his research. These levels are the only levels of the variable of interest to the researcher. In a psychological experiment, these levels might for example be a treatment and a control group. The purpose of the study then is to compare these two groups with each other. It is not of interest to compare these groups with other groups (unmeasured levels of the variable, for example other treatments) that were not included in the research. In a non-experimental setting, a variable with only a small subset of possible values might be treated as fixed effect, because all possible values can be included in the study (e.g. "gender", female and male). Furthermore, a researcher might be interested in generalizing the results to the levels of the variables that were included in the study. Imagine for example a survey study in which ten cities were selected at random. If the researcher does not want to generalize his results to all cities, but only to the ten cities that were included in the research, "city" can be treated as a fixed effect (Littel, Stroup, & Freund, 2002). Compared to the random effects model, the fixed effects model will probably produce smaller standard errors. Therefore it is more powerful. (Galwey, 2006).

When we speak of a random effect, we mean a statistical model which assumes that an independent variable is random. Such a model is generally used when it is assumed that the levels of the independent variable are a small subset of the possible values which one wants to generalize to. Therefore, a variable might be treated as random effect if we can think of the included levels of the variable as a sample of possible values of all possible levels of the variable. For example, in case of the survey study (as discussed in the last section), one would more naturally treat the effects of the variable "city" as random if the research's purpose is to generalize the results to the full population of cities. A disadvantage of the random effect model is that it will probably produce larger standard errors. Therefore, it is less powerful in comparison with the fixed effects model (Galwey, 2006).

An important difference between a fixed effects and a random effects model is the information the researcher is interested in. When a fixed effects analysis is conducted, the researcher is generally interested in explicitly comparing the scores on the dependent variable among the possible levels of the fixed variable. A researcher might for example want to compare the mean of a control group with the mean of different treatment groups. When a random effect analysis is conducted, the researcher is more interested in the degree of variance in the dependent variable that is explained by the random variable. So, in the case of a random effect, one is not interested in the means across the

different levels, but in the variance of means across the levels of a random factor (Littel et al., 2002).

To further illustrate the difference between fixed an random effects, imagine a study with a repeated measures design. In the repeated measures design, the same subjects are used with every condition of the research, for example a longitudinal study in which subjects receive a sequence of different treatments. Consider the following study as an example. Participants are randomly assigned to two different groups (treatment and control group), then their reaction time is measured in a number of different trials. By measuring their reaction time in multiple trials, it is possible to estimate not only the fixed effect (the effect of the different groups), but also the random effect (the effect of the subjects). Contrary to pure between-subjects design (in which we have an element of variance due to individual different that is combined in with treatment and error terms: SSTotal = SSTreatment + SEError), one is able to partition out the variability due to individual differences from treatment and error terms. The variability can be split up in between-treatments variability and within-treatments variability.

When someone wants to correct for the effect of subjects, it is therefore of crucial importance to choose a repeated measures design.

In a mixed-effects model, the random effect is modelled by a random intercept for every subject. The random effect itself is expressed in terms of variance. By doing so, the variance explained by the random effect can be compared to the variance as explained by the residuals. The comparison of these two variance terms make it then possible to determine the importance of the random effect in a given dataset. The underlying mechanism of a mixed-effects model will be explained more elaborately in the next section.

1.3.2 Mixed-effects modeling

In the following section, the idea of mixed-effects modeling is explained non-technically. A technical introduction of the concepts and formalism of the linear mixed effects models can be found in Appendix A.

Mixed-effects model is particularly useful in situations where repeated measurements are made on the same statistical units or where measurements are made on clusters of related statistical units. A mixed-effects model is a statistical model that contains both fixed effects and random effects (mixed effects thus). The random effect is modelled

as random intercept in a mixed-effects model (for further explanation see Appendix A and Section 1.3.3). It is also possible to include a random slope. As this however is not applicable for this thesis, this case will not be considered here. For readers who are interested in the application of a random slope, the article by Baayen, Davidson, and Bates (2008) is advised for further reading.

In a normal (multiple) regression analysis that is known to most researchers and commonly used in HCI studies, independence of observations is assumed. However, in a longitudinal study or when repeated measures on a subject are taken, these measurements tend to be positively correlated. Therefore, a regression analysis should not be the first-choice statistical analysis. In the design of a HCI usability, often repeated measures on subjects are taken, participants for example often have to evaluate several products by the means of the same scale (Hassenzahl & Monk, 2010, e.g.). As a consequence, these studies often face dependence of measurements (measurements of the same person are positively correlated). Nevertheless, these studies conduct regression analyses, ignoring the fact that these do not reflect the variance structure of the data appropriately. As the regression model treats all data as independent observations, degree of freedoms are overestimated and we face inflation of Type I error. This can be dangerous, because the parameter estimates and p values as gained by such an analysis may be biased. As in a mixed-effects model it is assumed that the observations are dependent, mixed-effects modeling can be an important tool for HCI researchers to solve this issue.

Baayen et al. (2008) discuss in their article a statistical approach known as *linear* mixed-effects modeling. These address drawbacks from traditional approaches to mixed-effects modeling. These drawbacks include:

- Deficiencies in statistical power
- The lack of a flexible method of dealing with missing data
- Disparate methods for treating continuous and categorical responses
- Unprincipled methods of modeling heteroscedasticity and non-spherical error variance

This approach will be used in the current study.

1.3.3 Mixed effects modeling and ARS

If someone would only analyse one judgement (one response to a Likert scale item), then the ARS would be part of ϵ, the general error in measurement. In this case, one cannot detect the ARS and/or account for it.

However, when repeated measures within participants are measured, it is possible to estimate the person specific response style bias. This can be done by using mixed-effects modeling. In mixed-effects modeling, the intercept for each subject can be modelled as random, which means that it differs for each person. By doing this, it can be controlled for the possible influence of the response style of a participant.

To make this more imaginable, data was simulated to give a simple example (the code of the data simulation can be found in Appendix B). Imagine the following research: In a survey research, it was tested whether the perceived attractiveness of a website influences its perceived usability. To test this, the researcher asked participants to rate the attractiveness and usability of four different websites. To measure these two variables, he used two 7-point Likert scale items ("I find the website usable" and "I find the website attractive").

The hypothesis was tested by correlating the mean of the rated attractiveness with the mean of the rated usability. This results in a correlation of e.g. r = 0.7. This correlation suggests a positive relationship between the two constructs. The researcher did however not account for a possible impact of the ARS.

As the participants had to rate the attractiveness and usability for more than one website, the data can be seen as repeated measures. There is more than one measure of a user's rating of the variables of interest. This extra information can be used to estimate the degree of a response style for each participant. Image that we do this, and find a ARS for one participant. Imagine further that this participant was a high ARS respondent. That means practically that the score for this participant is higher than the 'true' score, as it is influenced by the impact of the ARS. When we account for this response style, the score is adjusted (it is lowered) for each item. This results in a lower correlation, e.g. r = 0.02.

For a visual presentation of this situation, see Figure 1.2. The graphic represents the scores of the participant. On the x-axis the participant's scores for the four websites on attractiveness can be seen. On the y-axis his scores on the four websites on usability can be seen. In the graphic, the result of both situations is presented: In the first

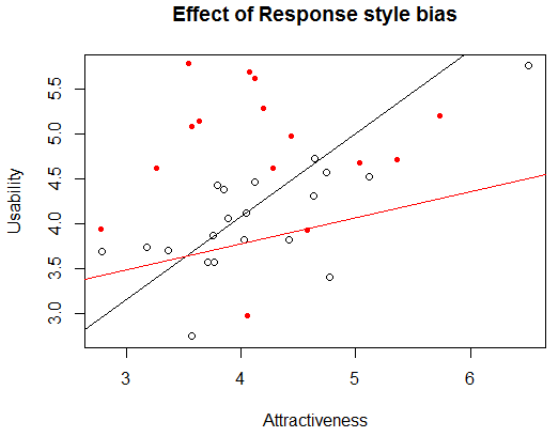

FIGURE 1.2: The ARS in action

situation, the data is not corrected, so the scores are quite high and result in a high correlation (see black line). In the second situation, the data is corrected for the influence of the ARS, that is, the score for each item is lowered. This results in a lower correlation (see red line).

In addition to the correction of a possible ARS, another advantage of the mixed-effects model approach is that the researcher can treat the items as well as the subjects as a random effect in the analysis. As the products used in a HCI study in general must be treated as a sample rather than an exhaustive list, it makes sense to treat items as a random effect as well. This makes it possible to conduct an analysis on the item level as well as on the subject level.

1.4 Aim of this thesis

The influence of the ARS is a threat for the validity of many HCI research results. Mixed effects modeling can be a promising tool for HCI researchers to deal with this. Therefore, in this thesis, we want to show that HCI studies are biased by the influence of the ARS and that mixed-effects modeling can be a reliable tool to account for it.

To make these aims concrete, the following research questions were tested:

1. *Does the impact of the ARS influence the validity of HCI research results?*
2. *Can this be dealt with appropriately by the means of mixed-effects modeling?*

These research questions were investigated by replicating a HCI study by Hassenzahl and Monk (2010).

The main objective of the study was to re-examine the relationship between beauty and usability. Their study focused on the interplay between four distinct constructs: Goodness (overall evaluation of a product in a given context), beauty (response to the Gestalt of the product), pragmatic quality (focuses on quality in use) and hedonic quality (focuses on personal needs). Four distinct data sets were used, which represented a wide range of websites, each of which sampled in a systematic manner. For set 1, participants had to rate 10 websites. For set 2, participants had to assess 60 website and for set 3, participants had to rate 30 websites. Set 4 consisted of ratings gathered via an on-line questionnaire. To measure pragmatic and hedonic quality, they constructed a short, eight-item version of the 21 item AttrakDiff2 (Hassenzahl, Burmester, & Koller, 2003). In a correlation analysis, they found that there is a significant relationship between beauty and pragmatic quality and beauty and hedonic quality for each data set. However, they suggested that any observed correlation for the relationship between beauty and pragmatic quality was mediated by goodness. A mediator analysis of the relationship between beauty, the overall evaluation and pragmatic quality indeed showed that the relationship between pragmatic quality and beauty was wholly mediated by goodness.

Following the recommendation of (Clark, 1973), Hassenzahl and Monk (2010) sampled the gained data in two different ways (aggregated over websites) and therefore considered a material and subject level. By doing so, they accounted for the ARS – by treating participants as a random variable in the sample that was aggregated over participants. However, Clark (1973) advised this procedure in order to account for the sampling error. Therefore, the impact of the ARS is not discussed in the article. Furthermore, they did not analyse their results by the means of a mixed-effects model. Therefore, they could not interpret the extent of the impact of the ARS (by comparing the variance as explained by the ARS to the varaince as explained by the residual errors).

In the current study, the same measures and materials were used to replicate the study by Hassenzahl and Monk (2010). After having replicated the research results of Hassenzahl and Monk (2010), it was first investigated whether the impact of the ARS is indeed a threat to the validity of HCI research results. After that, the data was analysed using mixed-effects modeling. By doing this, it was accounted for the influence of the ARS.

Additionally to this, a self constructed scale was used that is totally independent of the constructs measured in the to be replicated study. The purpose of this was to have

the 'ultimate' proof that the found relationships are an artefact of the response style bias in both constructs. Logically, the correlation of the measured constructs (beauty, goodness, pragmatic quality and hedonic quality) and a scale that is totally independent of these constructs should be equal to zero (as there is no relationship). However, when indeed an ARS can be found, this would result in a correlation between the independent scale and the rest of the scales.

Chapter 2

Method

In this study, the experiment by Hassenzahl and Monk (2010) was replicated. To be able to compare our results with the results of the original study, the same measures and materials were used and the procedure was similar to the procedure of the original study. In their study, Hassenzahl and Monk (2010) used four distinct data sets to gain information about the stability of the results by replication with different products and subjects. In this study, only one data set and sample was used.

2.1 Participants

72 students of the University of Twente participated either as partial fulfilment of course requirements or on voluntarily basis. The sample size was defined by the means of a simulation study (the R code can be found in Appendix C). Data of all 72 (47 female; age M = 21, SD = 3.8) participants were analysed. 53 participants of the sample were Dutch and 19 German. The experiment was approved by the faculty's ethics committee and all participants signed an informed consent before they participated in the experiment.

2.2 Measures

As in the original study, a short, eight-item version of the 21 item AttrakDiff questionnaire (Hassenzahl et al., 2003) was used to measure hedonic and pragmatic quality. A shorter test was required because of the large number of websites each participant had to evaluate. The same items as in the original study (Hassenzahl & Monk, 2010) were used. Four items measured hedonic quality and four items measured pragmatic quality. As in

the original study, the concepts beauty and goodness were measured with a single-item scale. Fro an overview of the items of these scales see Appendix D.

Additionally to these scales, one scale was added that was totally independent of the rest of the measured constructs. This scale was based on the 'attentiveness scale' of the Basic Positive Emotion construct of the expanded version of the Positive and Negative Affect Schedule (PANAS-X) test (D.Watson & Clark, 1994).

To get a scale that has the same construction as the rest of the scales, the items were implemented into a bipolar scale. The items of the Panas-X (D.Watson & Clark, 1994) were used but also the opposite of each item to get a bipolar item. In the following, this scale will be referred to as the "independent" scale. For an overview of all items see Appendix D.

The questionnaire was constructed on-line using LimeSurvey, an open source survey application (see `http://www.limesurvey.org)`, and hosted on the researcher's server. We made two different versions, a Dutch version and a German version. The only difference between these versions was the language; the design of the two different questionnaires stayed the same: At the top of the page, the participant could see the screen shot of the website. The width of this screen shot was the same for each website; the hight was adjusted by LimeSurvey automatically and was therefore not the same for each website. The size of the screen shot was more or less the same for each website (therefore, this could not influence participants) whereas the pictures were not dearranged. Below the screen shot the participants could see the eight-item version of the AttrakDiff questionnaire. To see an example of this design, see Appendix E.3. The independent scale was implemented in LimeSurvey in the same manner as the AttrakDiff scale (see Appendix E.4). Before participants could start with the evaluation and questionnaire of the first website, they were asked some demographical information (see Appendix E.2). Using a Python script the order in which the websites were presented to the participants was randomized (the Python script can be found in Appendix F).

2.3 Websites rated

Ten websites were selected randomly from the pool of websites used in the original study by Hassenzahl and Monk (2010). All websites have the primary aim of facilitating and supporting sales and on-line transactions to simulate real stores and travel agents. For an overview of all websites see Appendix G.

2.4 Procedure

All participants used an 1024 x 768 colour screen and a high bandwidth internet connection. The window of the browser was maximized to fit the whole computer screen. Each participant first had to reach the instructions on a website. The instructions required the participants to globally look at the screen shot of the website and then fill in the questionnaire regarding to their first impression (see Appendix E.1 for the full instructions). After they had read these instructions, the participants could start the experiment self paced by clicking on a button saying 'Start experiment'. When they finished the evaluation of the websites, participants had to answer the question 'How do you feel today?' and indicate their position on each item of the independent scale. As a cover story, the participants was told that in this study, also the possible relationship between usability evaluations of a website and the mood of participants was investigated.

2.5 Data analysis

The first step of the analysis was to investigate the intercorrelations of the measured constructs of different data aggregations (naive analysis, aggregated over websites and aggregated over subjects). Then, a linear regression (naive analysis) was conducted. This was followed by an analysis by the means of the mixed effects model approach, followed by a comparison of the naive analysis and the mixed effects model approach. As the unstable regression coefficients suggested that there was collinearity between the measured constructs, this was investigated. Next, to replicate the research results of the original study, a mediator analysis was performed with usability (pragmatic quality) as dependent variable, beauty as independent variable and goodness as mediator. As a series of simulations show that structural equation models perform better than multiple regression models when investigating a possible mediator effect (Iabucci, Saldanha, & Deng, 2007), this mediator analysis was performed by the means of structural equation modeling (SEM). To investigate the model fit of a SEM model, in most studies, the χ^2 test is used. However, this procedure has been criticised. First of all, the χ^2 test is sensitive to not only the sample size of a research, but also to possible violations of the multivariate normality assumption (Hu, Bentler, & Kano, 1992; West, Finch, & Curran, 1995; Curran, West, & Finch, 1996). Second, using the χ^2 as a measure of the model fit will lead to inflated type I error for model rejection (West et al., 1995; Curran

et al., 1996). In the statistical literature, there is consensus that one should avoid reporting all fit indices that have been developed since the early days of SEM. However, there is a certain disagreement on which fit indices out of the many available options should be considered for model evaluation (Schmelleh-Engel, Moosbrugger, & Mueller, 2003). Schmelleh-Engel et al. (2003) suggest to use the following model fit criteria, as they represent an adequate selection of indices which are frequently presented in current publications: χ^2 and its associated p value, the ratio of the χ^2 and its degree of freedoms (df), the Root Mean Square Error of Approximation (RMSEA) and its associated confidence interval (CI), the Standardized Root Mean Square Residual (SRMR), the Nonnormed Fit Index (NNFI) and the Comparative Fit index (CFI). In the current study, the χ^2 statistic and its associated p value, the ratio of the χ^2 and its df, the CFI, SRMR and RMSEA and its associated CI are presented. The data analysis is concluded on the psychometric level using a Principal component analysis (PCA), Explanatory Factor Analysis (EFA) & Confirmatory Factor Analysis (CFA).

Chapter 3

Results

The analysis was done using R, an open source language and environment for statistical computing (development core team, 2007), which is freely available at http://cran.r-project.org. For the required statistical methods for the mixed-effects modeling approach, the lme4 (Bates, 2005; Bates & Sarkar, 2007) package was used. The R code of the analysis can be found in Appendix H. All cases of the sample were analysed. After it turned out that the gained data was more complex than was thought beforehand, it was decided to conduct an exploratory analysis. Note that, as therefore different statistical tests were used, we face multiple testing which can result in an inflated Type I error (see e.g. Tabachnick & Fidell, 2005).

In the following, the results of the analysis will be presented. It will be referred to the measured constructs as following: B (Beauty), G (Goodness), H (Hedonic quality), P (Pragmatic quality) and I (Independent scale).

3.1 Intercorrelations

3.1.1 Correlations as gained by naive analysis

First of all, the correlations of the different constructs as conducted by a naive analysis (not corrected for the effect of the ARS or the effect of the websites as a common factor) were calculated. Note that this analysis ignores the fact that the measurements are not independent and might therefore be biased.

You can see all correlations in Table 3.1. A visualisation of these results is given in Figure 3.1.

TABLE 3.1: Correlations of the constructs (naive analysis)

	B	G	H	P
B	1.00	0.71	0.82	0.33
G	0.71	1.00	0.70	0.59
H	0.82	0.70	1.00	0.28
P	0.33	0.59	0.28	1.00

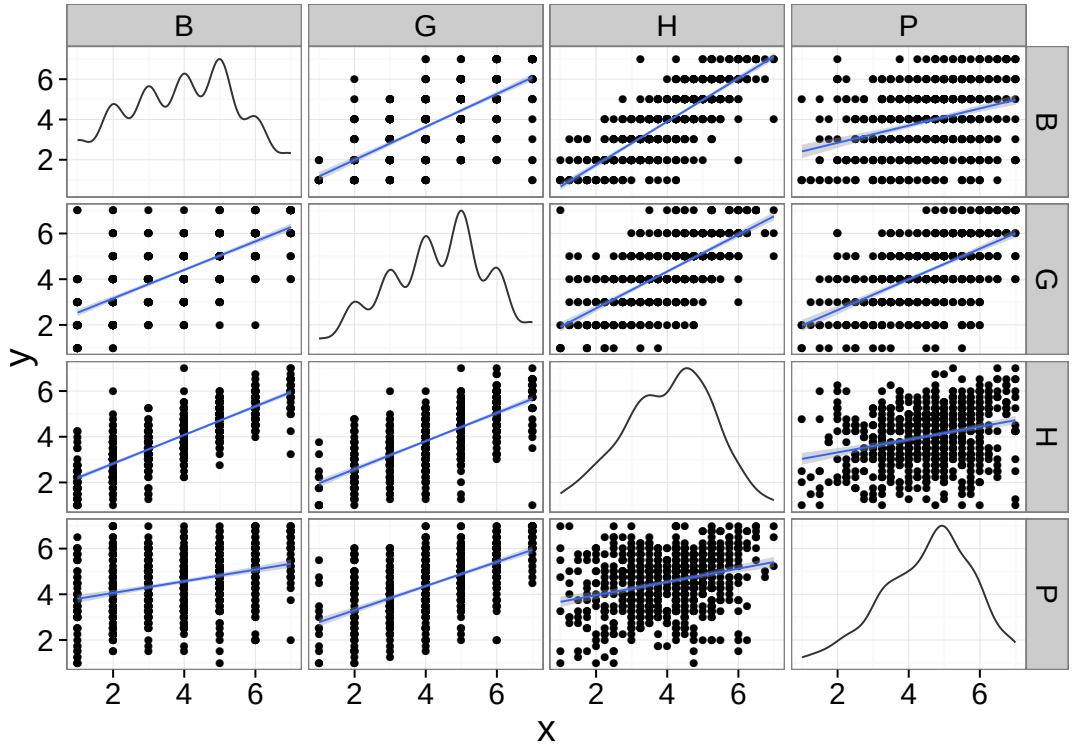

FIGURE 3.1: Correlations as gained by naive analysis

We can see that most correlations are quite high. Correlations with the construct pragmatic quality were however only moderate.

3.1.2 Correlations when data was aggregated over websites

To have a look at the effect of the ARS, the data was aggregated over the websites. By doing this, the effect of the websites was removed and what was left over was the effect of the ARS. The correlations are shown in Table 3.2, a presentation is given in Figure 3.2.

We can see that the correlations are still quite high. Correlations with the independent scale were however quite small. This suggests that the influence of the ARS

TABLE 3.2: Correlations of the constructs (when aggregated over websites)

	B	G	H	P	I
B	1.00	0.65	0.81	0.22	0.27
G	0.65	1.00	0.68	0.49	0.26
H	0.81	0.68	1.00	0.26	0.20
P	0.22	0.49	0.26	1.00	0.02
I	0.27	0.26	0.20	0.02	1.00

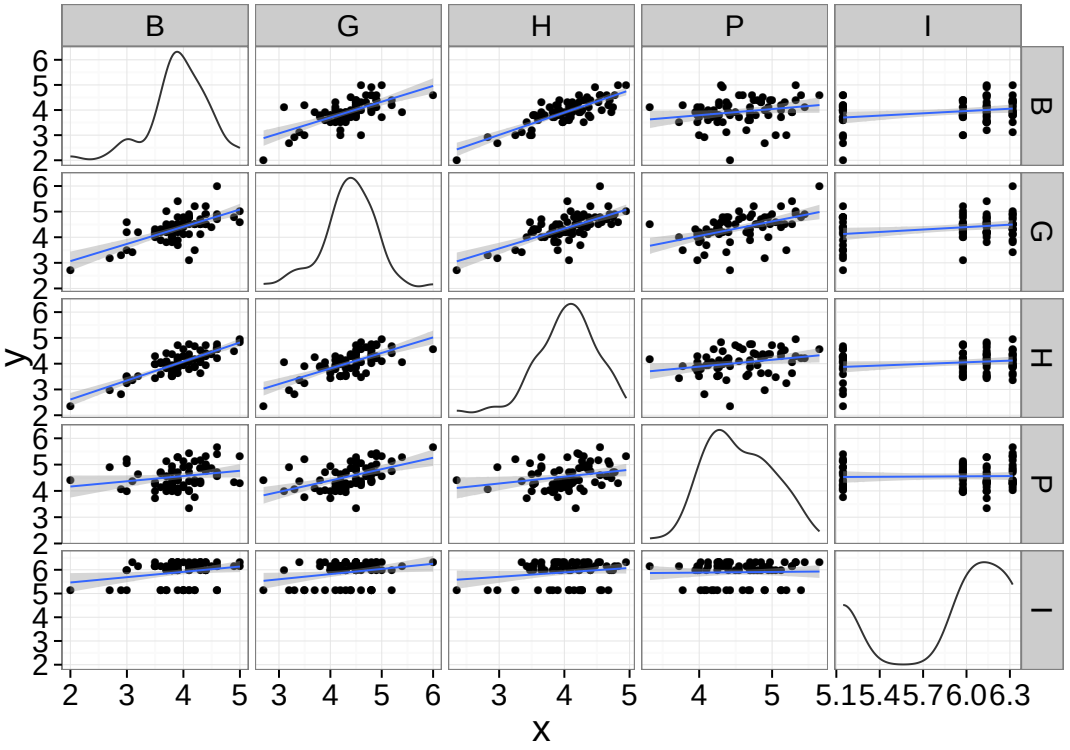

FIGURE 3.2: Correlations as gained when data was aggregated over websites

was very small. Furthermore, the correlations with the construct pragmatic quality were quite low.

3.1.3 Correlations when data was aggregated over subjects

Next, the data was aggregated over subjects. By doing this, the effect of the subjects (and therefore the effect of the ARS) was removed and what was left over was the effect of the websites as a common factor. All correlations are shown in Table 3.3, a visualization is given in Figure 3.3.

You can see that these correlations are quite high for all measured constructs.

TABLE 3.3: Correlations of the constructs (when aggregated over subjects)

	B	G	H	P
B	1.00	0.94	0.98	0.71
G	0.94	1.00	0.94	0.83
H	0.98	0.94	1.00	0.70
P	0.71	0.83	0.70	1.00

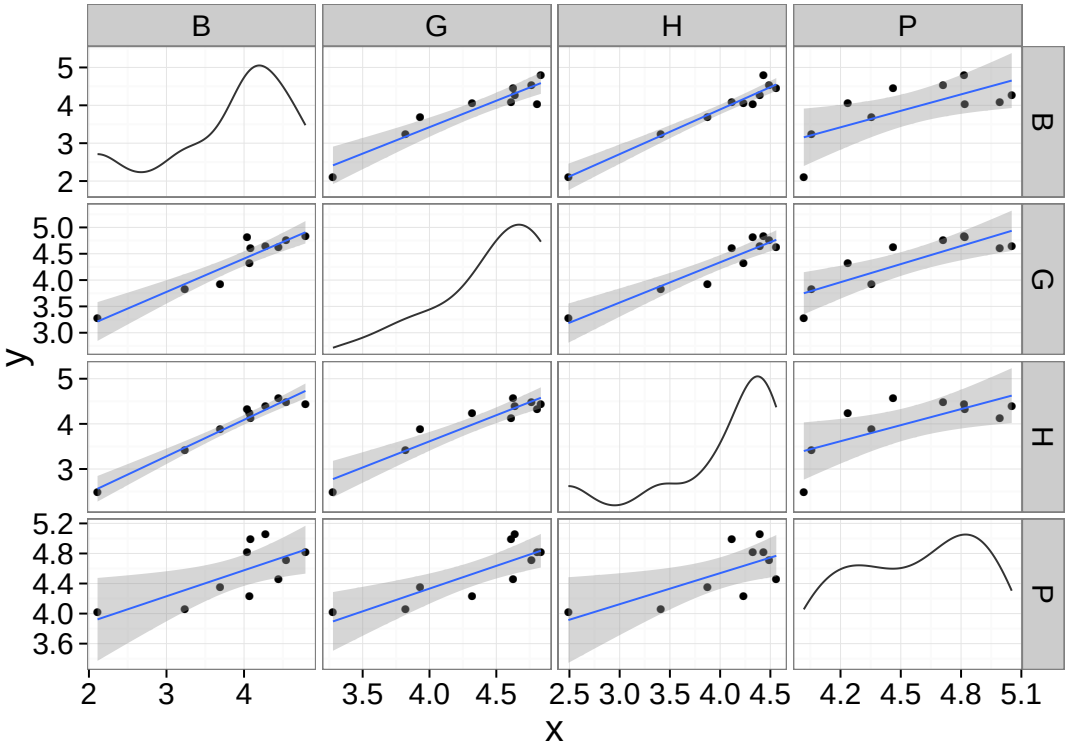

FIGURE 3.3: Correlations as gained when data was aggregated over subjects

3.2 Normal regression (naive analysis)

First of all, the relationship between the constructs pragmatic quality, beauty, goodness and hedonic was tested by a "normal" linear regression (naive analysis) to be able to compare the results with the results of the mixed-effects approach. Again, note that the naive analysis assumes that there is no correlation within the subjects and websites which is no the case here and that the estimates as gained by the naive analysis might therefore be biased.

To test beforehand whether the websites as a common factor explained any variance at all, a simple linear regression model with as outcome variable pragmatic quality

and as independent variables hedonic quality, goodness, beauty and websites was conducted. Automated model comparison showed that the websites as common factor did not contribute to the explained variance. Therefore, we omitted this factor from further analysis.

To choose the best model to explain the relationship between the constructs different models were compared with each other. The corrected R square and the akaike information criterion (AIC) was calculated for each possible model. We did not consider any interaction effects. It could be concluded that the model with as independent variables goodness and hedonic quality was the best model as based on the highest R square and the lowest ACI. Furthermore, the model comparison showed that the parameter estimates were quite unstable. This suggests that there was collinearity between the measured constructs. This suggestion is tested in Section 3.5. An overview of the parameter estimates as gained by the model comparison can be found in Appendix I.

The model fit of the statistical model as measured by the AIC was 2011.04.

The coefficients and model effects are provided in Table 3.4. The influence was significant for the goodness construct (T (717) = 19.9, $p < 0.001$) as well as for the hedonic quality construct (T (717) = -6.42, $p < 0.001$). Compared to the intercept, the effect sizes of these constructs were however only small.

TABLE 3.4: Estimated parameter coefficients and model effects

	Estimate	Std Error	t value	p value
Intercept	2.61	0.13	19.9	< 0.001
G	0.69	0.04	19.00	< 0.001
H	0.27	0.04	-6.42	< 0.001

3.3 Mixed-effects models

The relationship between pragmatic quality and beauty, goodness and hedonic quality was tested with mixed-effects models as well. This was done on two different levels:

1. The subject level (Random effect is ARS)

2. The material level (Random effect is common factor of the websites)

TABLE 3.5: Coefficients and model effects for fixed effects (subject level)

	Estimate	Std Error	t value
Intercept	2.59	0.14	19.30
G	0.71	0.04	19.54
H	0.28	0.04	6.74

TABLE 3.6: Variance and SD for the random effect and residual (subject level)

	Variance	Std Dev
Subjects (Intercept)	0.01	0.30
Residual	0.86	0.93

3.3.1 The subject level

To test the influence of the ARS as random effect, first of all we compared different models in which the subjects variable was defined as random effect.

The model comparison showed that the best model on the subject level was a mixed-effects model with as predictors hedonic quality and goodness. This was concluded on the basis of the lowest AIC.

The model was fitted using restricted maximum likelihood estimation (REML), a modification of maximum likelihood estimation that is more precise for mixed-effects modeling. The purpose of maximum likelihood estimation is to find those parameter values that, given the data and the choice of the model, make the model' predicted values most similar to the observed values. As measured by the AIC, the model fit was 2009.

Coefficients and model effects as estimated for the fixed effects can be found in Table 3.5 and variance and standard deviation of the random effect and residual in Table 3.6. It can be seen that the parameter estimate of the intercept was smaller compared to the naive analysis whereas the parameter estimates of goodness and hedonic quality were bigger. These differences were however only small. Furthermore, the variance of the random effect was very small, especially when compared to the variance of the residual. This will be discussed further in Section 3.4.

3.3.2 The material level

As in 3.3.1, a model comparison was performed first. The model comparison showed that the best model on the material level was a mixed effects model with as predictors hedonic quality and goodness. This was concluded on the basis of the lowest AIC.

The same model fitting procedure was handled as explained in 3.3.1. The model fit as measured by the AIC was 2013.

The coefficients and model effects as estimated for the fixed effects can be seen in Table 3.7 and the variance and standard deviation for the random effect and residual in Table 3.8. Compared to the naive analysis and the other mixed-effects model, the parameter estimates (with exception of the estimate for the goodness construct) were bigger in this analysis. The variance of the random effect was again quite small, but bigger than the variance for the random effect on the subject level.

TABLE 3.7: Coefficients and model effects for fixed effects (material level)

	Estimate	Std Error	t value
Intercept	2.76	0.15	17.84
G	0.68	0.04	19.00
H	0.30	0.04	-6.81

TABLE 3.8: Variance and SD for the random effect and residual (material level)

	Variance	Std Dev
Material (Intercept)	0.04	0.21
Residual	0.91	0.95

As for the naive analysis, it could be found for both mixed-effects models (subject as well as material level) that the parameter estimates were quite unstable. You can find the estimated parameter estimates in Appendix I.

As it is mathematically difficult to estimate the degree of freedoms in a mixed-effects modeling (Baayen et al., 2008), p values for the parameters as estimated by the two mixed-effects models cannot be given here. The only way to yet derive p values would require the use of Bayesian statistics, which goes beyond the scope of this thesis.

3.4 Comparison of the naive model and the mixed-effects model

To investigate whether the mixed-effects model that corrects for the ARS (subject level) fits the data better than the model of the naive analysis, these two models were compared with each other. When comparing the parameter estimates of the different models, it can be seen that the naive analysis resulted in lower estimates for the effect of goodness and hedonic quality and a higher estimate for the intercept compared to the mixed-effects model (subject level). These differences are however only small and non significant.

The R^2 and the mean squared error of prediction (MSEP) was calculated for both models. As the value of the R^2 increases when a model has more predictors, we also provide the adjusted R^2 which increases only if the new term improves the model more than would be expected by chance. You can see the results in Table 3.9. The higher R^2 value of the mixed-effects model suggests that this is the superior model. The smaller value of the adjusted R^2 value suggests that this high value can be associated with the additional estimator in the mixed effect model (the influence of the subjects as random effect). Although the adjusted R^2 value of the mixed-effects model is still higher than the adjusted R^2 value of the naive model, the difference is very small. The MSEP of the mixed-effects model is lower which suggests that the estimation of this model is more precise.

TABLE 3.9: Comparison of the naive and the mixed-effects model

Model	R^2	adjusted R^2	MSEP
Naive model	0.39	0.39	0.94
Mixed-effects model	0.47	0.41	0.81

Although this suggests that the mixed-effects model is the better model, the results of the mixed model show that the variance of the random effect is very small (0.01). This suggests that the random effect does not contribute to the explained variance of the data. This becomes particularly evident when comparing the variance of the random effect with the variance of the residuals (0.68). This value is 68 times higher than the value of the random effect. When plotting the range of the random intercepts of the mixed-effects model next to the range of the residuals of the naive model (see Figure 3.4), it can be seen graphically that the random effect does not explain much

variance compared to the residuals of the naive model. This effect becomes even more evident when both histograms are centered to zero (see Figure 3.5. We conclude that the mixed-effects model (subject level) is not superior to the naive model without random effect.

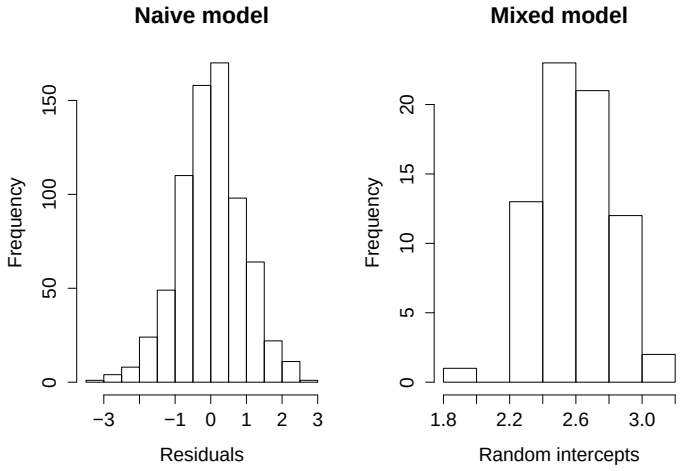

FIGURE 3.4: Histograms of the naive model and the mixed-effects model

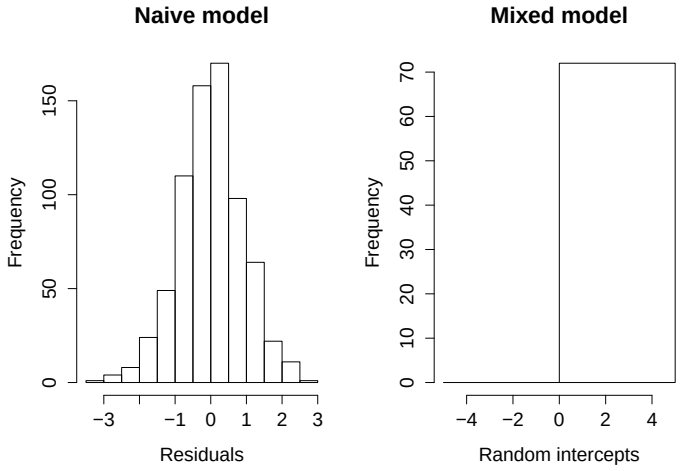

FIGURE 3.5: Histograms of the naive model and the mixed-effects model when both are centered to zero

3.5 Collinearity

The instability of the parameter estimates (see Appendix I) suggested that there might be collinearity between the measured constructs. To test this, the variance inflation factor (VFI) was calculated for each $\hat{\beta}_i$. See Table 3.10 for the results.

TABLE 3.10: VFI for each predictor

	VFI
B	147.85
G	44.21
H	49.91

A common rule of thumb is that if the value of the VFI is bigger than five, then collinearity is high. Ten has also been proposed as cut off value (see Kutner, Nachtsheim, & Neter, 2004). In this case, it does not matter whether a conservative or less conservative cut off value is chosen. It should be clear that we face multicollinearity here.

3.6 Mediator analysis

As in common SEM notation (Kline, 2010), the structural model and the measurement model are provided in one model. You can find this model in Figure 3.6. Pragmatic quality was the latent variable of the model. For the sake of simplicity, the variables are indexed with a G, P and B, representing the different constructs. For the mediation, the same variables were chosen as in the original study. Pragmatic quality was the dependent variable, beauty the independent variable and goodness the mediator variable. As in "mediator" terms, path a represents the indirect effect, path b the direct effect and path c the total effect. As the constructs of beauty and goodness were single-item measurements only, the error term influences these constructs directly (which is not applicable to the beauty construct, as it is an exogenous construct in this model). It has to be noted here, however, that in practice, it is not possible to estimate the amount measurement error of single-item measurements. This issue will be discussed in Section 4.7.

The results of the SEM show that the effect of beauty on pragmatic quality was not significant (Z = -1.4, p = 0.15). The effect of beauty on goodness (Z = 7.2, p = <

0.01) as well as the effect of goodness on pragmatic quality (Z = 4.50, $p = < 0.01$) was highly significant. This suggests that the relationship between pragmatic quality and beauty was fully mediated by goodness.

In all graphics, the dominant symbolic language in the SEM world is used. A list of the used symbols can be found on page 5.

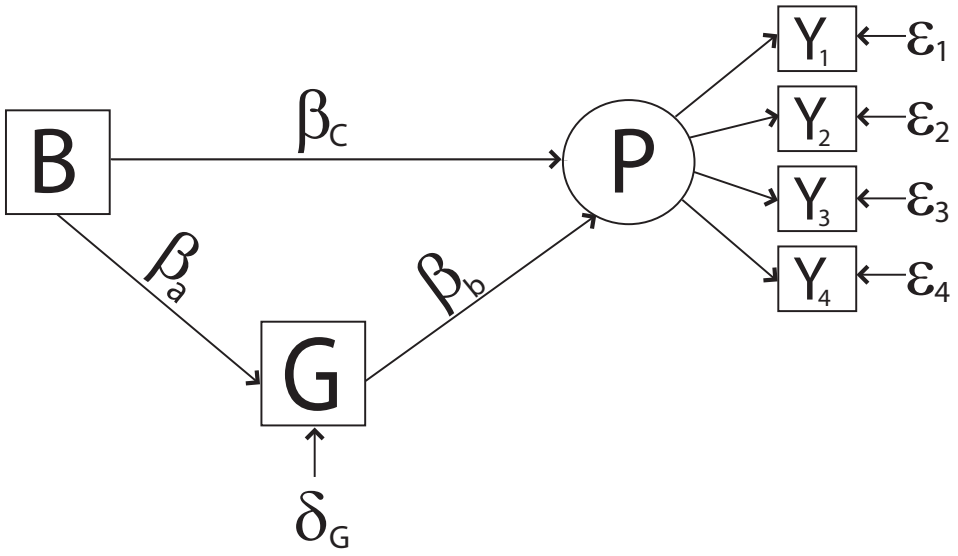

FIGURE 3.6: Mediation: Structural & measurement model

3.6.1 Model fit

You can find the measures of the χ^2 statistic and the associated p value, $\frac{\chi^2}{df}$, the CFI, SRMR and RMSEA and its associated CI in Table 3.11. For an explanation of the choice of these model of fit measures see Section 2.5.

TABLE 3.11: Measures of model fit

χ^2	p value	$\frac{\chi^2}{df}$	CFI	SRMR	RMSEA	RSMEA CI lower	RSMEA CI upper
18.15	0.02	2.27	0.96	0.04	0.13	0.05	0.21

The underlying approximation of the measurement of model fit is different for the different model fit measures. Therefore, the results have to be interpreted differently and we have to discuss each model fit measurement separately. The χ^2 test statistic

is used for hypothesis testing to evaluate the appropriateness of a structural equation model. The null hypothesis states that the proposed model fits the data well. Therefore, accepting the null hypothesis supports the researcher's belief. This makes the χ^2 statistic a "badness of fit" index; a large χ^2 statistic indicates that the model was poorly fitted. In this case, the χ^2 value is quite high and the p value is below 0.05 ($p = 0.02$). This means that the null hypothesis has to be rejected. We can therefore conclude that the model does not fit very well. The second model fit measure is the χ^2 divided by the degree of freedoms (df) of the proposed model. There is no universally agreed upon standard as to what is a good and bad fitting model as indicated by the second measure. Iacobucci (2009) advice considering the model fit good when the value is about three or below three. In our case, the value is below three which suggests that the model fits well. The third measure, the CFI, is an incremental fit index. An incremental fit index is analogous to R^2. Therefore, a value of zero indicates having the worst possible model and a value of one indicates having the best possible model. A rule of thumb is that the value should be at least 0.95 (Schmelleh-Engel et al., 2003; Iacobucci, 2009). In our case, the values is bigger than 0.95, which suggests that there is a good model fit. The SRMR is a residual based index. The underlying assumption here is that when the model fits well, the residuals (differences between the model implied covariance matrix and the sample covariance matrix) should be small. Schmelleh-Engel et al. (2003) suggest that the value of this measure should be smaller than 0.05 to be considered as well-fitting model. This is the case which confirms the suggestion that our model is well fitting. The last measure we used is the RMSEA. Like the SRMR, it represents a residual based index and its value should be less than 0.05 (Schmelleh-Engel et al., 2003). More specifically, MacCallum, Browne, and Sugawara (1996) suggest that 0.01, 0.05 and 0.08 indicate excellent, good and medicore fit respectively. However, in our case, the value is bigger than 0.08 which suggests that the model fit is not as good as the other model fit methods suggest. For a better interpretation of the value (Schmelleh-Engel et al., 2003), the 90% CI was calculated. Ideally, the lower bound of the CI should include or be very near to zero and should not be bigger than 0.05 (Schmelleh-Engel et al., 2003).

With the exception of two model fit measurements, all gained values suggest that the model fits well. The two model measurements that point into a different direction are the χ^2 test statistic and the RMSEA. We already pointed out that the χ^2 test statistic is not reliable. In this case, it seems as if the RMSEA is not very reliable as well. Although its CI is not bigger than 0.05 (as suggested by the rule of thumb

for a well fitting model), it is quite big, which suggests that the measure is not very precise. Furthermore, there is a greater sampling error for a model with small df and low N, especially for the former. Therefore, models with small df and/or low N can have artificially large values for the RMSEA. Kenny, Kaniskan, and McCoach (2011) argue for this reason to not even compute the RMSEA for low df models. As the df and sample size of our model is not very big, the RMSEA is probably biased and the other model fit measurements should be trusted. Therefore, it can be concluded that the model fit is quite good.

3.7 Psychometric level

As could be seen, we face multicollinearity here. This suggests that there is one underlying variable for all constructs. To find this underlying variable or at least get more insight in the structure, we switched to the psychometric level and further investigated the results by the means of psychometric research methods. For the psychometric analysis, the data was collapsed over websites. Therefore, any effect that could be found represents the personal preferences of subjects.

To possibly reduce the correlated observed variables to a smaller set of important independent composite variables, the first step was to conduct a principal component analysis. The results suggest that the construct beauty is indistinguishable of the construct hedonic quality. Goodness is more strongly related to hedonic quality and beauty, but has also some link to pragmatic quality.

The results of the PCA can be found in Appendix J. To confirm these suggestions, an exploratory factor analysis was conducted to explore the possible underlying factor structure of the tested constructs. The results of the EFA can be found in Appendix J as well.

3.7.1 Confirmatory factor analysis

The principal component analysis as well as the exploratory factor analysis suggest that beauty and hedonic quality are indistinguishable. This suggests that the model of now consisting five factors (beauty, goodness, hedonic quality, independent scale (ARS) and pragmatic quality) could be reduced to a four factors model (beauty + hedonic quality, goodness, independent scale (ARS) and pragmatic quality). Therefore, the next step was to compare the five factor model with the four factors model. To do this, two

confirmatory factor analyses were conducted. The first had five factors, the second four. Afterwards, the akaike information criterion of the two models were compared to see which model fits the data better.

3.7.1.1 Five factors model

The measurement model of the five factors model can be seen in Figure 3.7. As the constructs beauty and goodness were measured by the means of a single item measure, it was not possible to estimate their factor loadings. This will be explained and discussed in Section 4.7.

The results of the five factor model can be seen in Table 3.12, Table 3.13 and Table 3.14. It can be seen that goodness is more strongly related to hedonic quality and beauty but has also some link to pragmatic quality.

TABLE 3.12: Five factors model: Estimates, std error, z value and p value

Latent variables	Estimate	Std.error	Z-value	p value
Pragmatism =				
PQ1	1.000			
PQ2	1.100	0.140	7.875	0.000
PQ3	1.069	0.155	6.908	0.000
PQ4	1.243	0.156	7.982	0.000
Hedonism =				
HQ1	1.000			
HQ2	0.949	0.104	9.161	0.000
HQ3	1.020	0.103	9.912	0.000
HQ4	0.695	0.095	7.308	0.000
ARS =				
indep1	1.000			
indep2	0.935	0.103	9.097	0.000
indep3	0.964	0.121	7.962	0.000
indep4	0.535	0.130	4.123	0.000
B =				
Beauty	1.000			
G =				
Goodness	1.000			

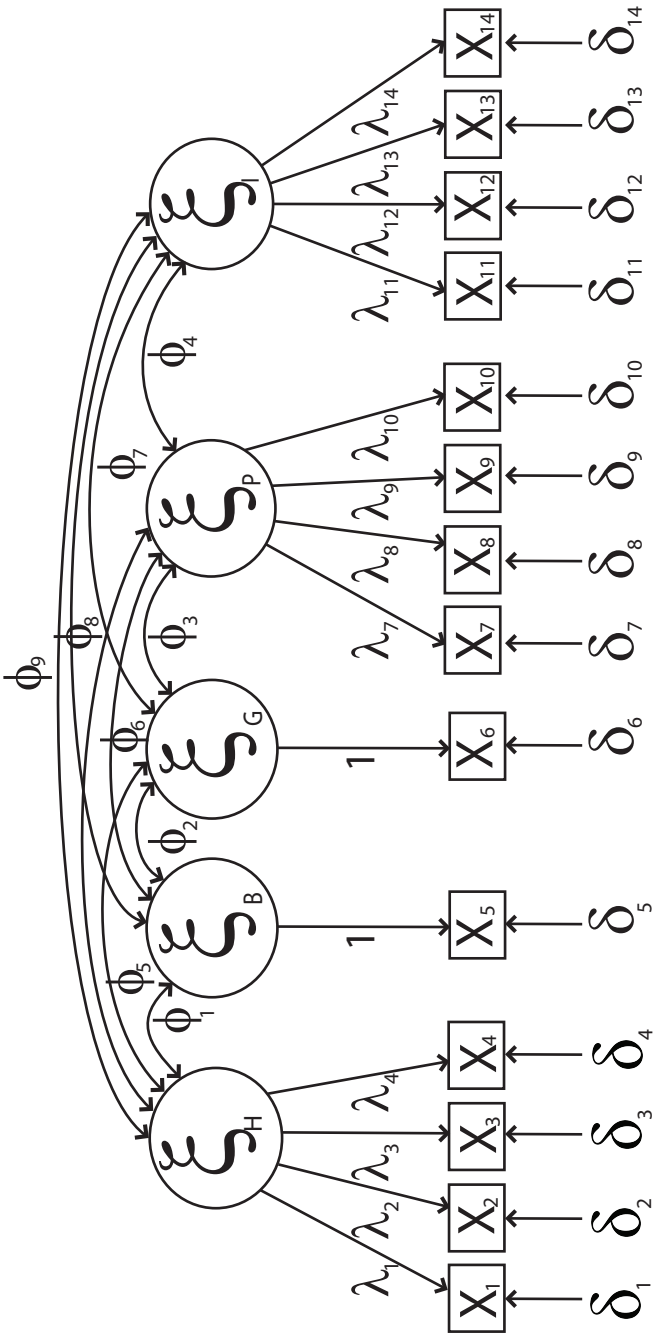

FIGURE 3.7: Measurement model five factors model

Table 3.13: Five factor model: Covariances

Latent variables	Estimate	Std.error	Z-value	p value
Pragmatism				
Hedonism	0.062	0.028	2.216	0.027
ARS	0.149	0.065	2.279	0.023
B	0.052	0.027	1.915	0.056
G	0.121	0.033	3.701	0.000
Hedonism				
ARS	0.191	0.079	2.421	0.015
B	0.220	0.043	5.110	0.000
G	0.189	0.041	4.583	0.000
ARS				
B	0.150	0.077	1.948	0.051
G	0.194	0.081	2.385	0.017
B				
G	0.184	0.040	4.628	0.000

Table 3.14: Five factor model: Variances

Estimate	Variance	Std error
PQ1	0.113	0.022
PQ2	0.061	0.015
PQ3	0.120	0.024
PQ4	0.070	0.018
HQ1	0.074	0.017
HQ2	0.100	0.020
HQ3	0.085	0.019
HQ4	0.108	0.020
indep1	0.677	0.143
indep2	0.136	0.076
indep3	0.626	0.132
indep4	1.285	0.219
Beauty	0.000	
Goodness	0.000	
Pragmatism	0.164	0.044
Hedonism	0.244	0.053
ARS	1.323	0.325
B	0.274	0.046
G	0.292	0.049

3.7.1.2 Four factors model

The results of the four factors model can be found in Table 3.15, Table 3.16 and Table 3.17. The measurement model of the four factors model is shown in Figure 3.8. Again, the factor loading of the construct goodness could not be estimated which will be explained elaborately in Section 4.7.

TABLE 3.15: Four factors model: Estimates, std error, z value and p value

Latent variables	Estimate	Std.error	Z-value	p value
Pragmatism =				
PQ1	1.000			
PQ2	1.094	0.139	7.893	0.000
PQ3	1.065	0.154	6.927	0.000
PQ4	1.240	0.155	8.027	0.000
Hedonism =				
HQ1	1.000			
HQ2	0.950	0.104	9.094	0.000
HQ3	1.019	0.104	9.803	0.000
HQ4	0.700	0.095	7.330	0.000
Beauty	0.912	0.094	9.716	0.000
G =				
Goodness	1.000			
ARS =				
indep1	1.000			
indep2	0.933	0.103	9.095	0.000
indep3	0.964	0.121	7.972	0.000
indep4	0.535	0.130	4.123	0.000

Comparison of the two models shows that we should retain the four factors model. The akaike information criterion of the model with four factors only is smaller than the akaike information criterion of the five factor model, which suggests that the model with four factors fits the data better (AIC for the model with five factors 1538.5, for the model with four factors = 1534.2).

In the results of four factors model we can see that variance of the independent scale ("ARS") was quite high compared to the rest of the scales.

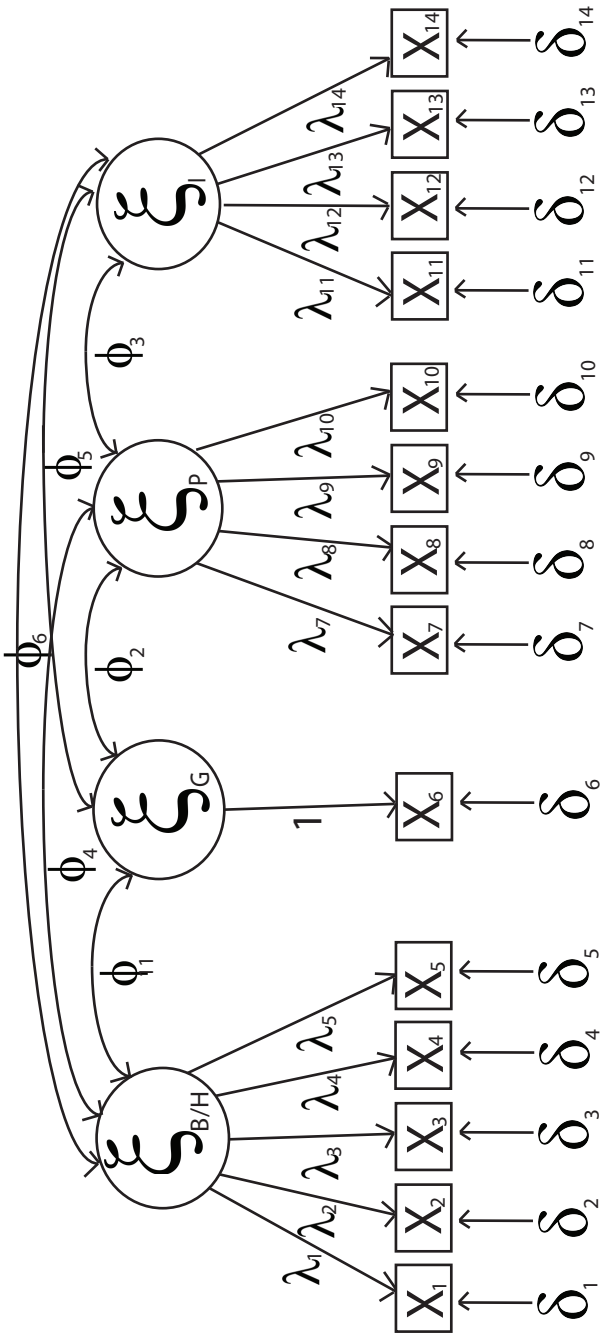

FIGURE 3.8: Measurement model four factors model

TABLE 3.16: Four factor model: Covariances

Latent variables	Estimate	Std.error	Z-value	p value
Pragmatism				
Hedonism	0.061	0.028	2.210	0.027
G	0.121	0.033	3.701	0.000
ARS	0.149	0.066	2.281	0.023
Hedonism				
G	0.192	0.041	4.644	0.000
ARS	0.185	0.078	2.375	0.018
G				
ASR	0.194	0.081	2.384	0.017

TABLE 3.17: Four factor model: Variances

Estimate	Variance	Std error
PQ1	0.112	0.022
PQ2	0.062	0.015
PQ3	0.121	0.024
PQ4	0.069	0.018
HQ1	0.076	0.017
HQ2	0.101	0.020
HQ3	0.087	0.019
HQ4	0.107	0.019
Beauty	0.072	0.015
Goodness	0.000	
indep1	0.676	0.142
indep2	0.138	0.076
indep3	0.624	0.132
indep4	1.284	0.219
Pragmatism	0.165	0.044
Hedonism	0.242	0.053
G	0.292	0.049
ARS	1.324	0.326

3.7.1.3 Model fit of the four factors model

The same measures were used as for the measurement of the model fit of the SEM. For an explanation of the underlying approximations and interpretations, see Section 2.5 and Section 3.6.1.

The results of all model fit measurements suggest that the model fit is moderate. Again, the results of the χ^2 test statistic and its associated p value point into another direction. However, because of aforementioned reasons (see Section 3.6.1), these results can be neglected.

TABLE 3.18: Four factors model: Measures of model fit

χ^2	p value	$\frac{\chi^2}{df}$	CFI	SRMR	RMSEA	RSMEA CI lower	RSMEA CI upper
109.1	¡0.01	1.51	0.94	0.06	0.08	0.05	0.06

Chapter 4

Discussion

In this chapter, we will discuss the results and answer the research questions of this study.

4.1 Acquiescence response style bias

One research question of this study was whether the validity of HCI studies is threatened by the impact of the acquiescence response style (ARS). We hypothesized that the high correlations as found in many HCI studies that investigate the relationship between multiple constructs (e.g. between usability and beauty) can be attributed to the influence of the ARS and do not represent a true relationship. To be able to answer this research question, the study by Hassenzahl and Monk (2010) was replicated. The main objective of their study was to re-examine the relationship between beauty and usability. The study focused on the interplay between four distinct constructs: Goodness, beauty, pragmatic quality and hedonic quality. To measure pragmatic and hedonic quality, they constructed a short, eight-item version of the 21 item AttrakDiff2 (Hassenzahl et al., 2003). In a correlation analysis, they found that there is a significant relationship between beauty and pragmatic quality and beauty and hedonic quality.

In this study, the same measures and materials were used to replicate the study. When we had a closer look at the correlations between the different constructs when the data was aggregated over websites, it seemed at first sight as if there was indeed an effect of the ARS. It could be seen that there were quite high correlations between the measured constructs. However, the correlations with the independent scale which was totally independent of the rest of the constructs were very small. Furthermore, the

correlations were still high when the data was pooled or aggregated over participants. This suggests that the high correlations cannot be attributed to the ARS but rather indicate that a participant who rated the website high on one of the scale also rated it high on the other scales and vice versa. It can be concluded that the effect of the ARS is so small that it can be neglected.

Based on earlier research in other research fields than HCI, we expected beforehand that without the influence of the ARS, the relationship of the constructs would be smaller or in fact totally fade away. Research results suggested that estimates of the relationship between observed variables can be misleading when it is not accounted for the possible impact of response styles (see e.g. Baumgartner & Steenkamp, 2001). As there is no literature about the impact of response styles in HCI research, knowledge is limited and we could only make assumptions. However, our review of methodology and research results of some typical HCI research disciplines suggested that these are quite vulnerable for the ARS.

Contrary to the results gained in the marketing and consumer research field (see e.g. B.Weijters, 2006), it can be concluded that the influence of the ARS can be neglected in the current study. It has to be noted here that this conclusion is however limited to the current research design. Practically this means that this conclusion can be drawn for the current sample and used stimuli only. As the current research sample consisted of mainly psychology students, the sample was not very diverse. Although the sample consisted of Dutch and German students, the participants all came from the same culture (European). As several studies have documented cross-cultural differences in ARS (e.g. Cunningham et al., 1977; England & Harpaz, 1983; Morris & Pavett, 1992; Riordan & Vandenberg, 1994), it is possible that although the current sample (Dutch and German people) did not show a tendency for an ARS, it is possible that this tendency can be found for other culture. As research shows that Asian people generally show a higher tendency for the ARS (e.g. Riordan & Vandenberg, 1994), it could be that in an Asian sample, the research results validity is in fact threatened by the influence of the ARS. To find out whether the results of the current study can be generalized, it is important to replicate this study with different samples and stimuli. Furthermore, there are other HCI research fields sensitive for the influence of the ARS. As mentioned Section 1.2.5, research that makes use of the TAM is prone to the influence of the ARS. We conclude that more research is needed to investigate the impact of the ARS in other HCI research fields.

The notion that the ARS is not a thread for the current research setup is in line with the paper "The great response style bias" by Rorer (1965). Rorer (1965) points out that various measures of the ARS generally fail to intercorrelate. He concludes that although such a thing was possible there had in fact been no demonstration that the acquiescent responding to scales independent of meaning had ever been of anything but the most trivial importance.

It has to be noted that the current study focused on the effect of the ARS only – assuming that this research area is most vulnerable for this particular response style. We can not rule out the possibility that the research results are threatened by the influence of other response styles.

The results of further analyses, in particular of the mixed effects model, confirm the notion that the influence of the ARS is so small that it can be neglected. These results will be discussed in the next section.

4.2 Mixed effects models

To be able to interpret the results of our study in the light of the influence of the subjects level (influence of the ARS) and on the material level (websites as common factor), we used two different mixed effects models. In the first model, the subjects were defined as random effect and in the second model, the websites as common factor were defined as a random effect.

4.2.1 Subject level

In the first model, we defined the influence of the participants as a random effect and therefore accounted for the influence of the ARS. By doing this, the intercept was adjusted for each subject. Hereby the influence of the ASR was removed. However, the variance of the random effect was very small.

In the perspective of a mixed effects model, this can be interpreted in two different ways:

1. Subjects filled in the questionnaire randomly, producing small correlations
2. The RS had no influence on the results

We could see that the results as gained by the naive analysis are similar to the results as gained by the mixed effects model. Therefore, in this case it is more intuitive

to accept the second way of interpreting the small variance of the random effect: The effects of the ARS were so small that its influence can be neglected.

4.2.2 Material level

The model in which the websites as common factor were defined as a random effect showed that the influence of the different websites were bigger than the influence of the ARS, but however not very big as well. Again, this can be interpreted in two different ways:

1. Subjects filled in the questionnaire randomly, producing small correlations
2. The websites were too similar. Therefore, participants could not distinguish between the different websites very well

When investigating the ratings for the different websites these ratings are quite similar. There is no website that stood out. In this case it is more reasonable to accept the second way of interpreting the small variance of the random effect: Since the websites were so similar that participants could not distinguish them very well they responded similarly to these different stimuli.

4.3 Mediator analysis

In a mediator analysis, Hassenzahl and Monk (2010) found that there was a significant relationship between beauty and pragmatic quality and beauty and hedonic quality and showed that any observed correlation for the relationship between beauty and pragmatic quality was mediated by goodness. In this study we replicated these results with a mediator analysis with the same constructs. As literature shows that analysis of a structural equation model is superior to multiple regressions to investigate a possible mediator effect, we used a structural equation model. We arrived at the same conclusions as Hassenzahl and Monk (2010). The relationship between pragmatic quality and beauty was fully mediated by goodness. Furthermore, different measures of model fit of the mediator model showed that the model fits the data well.

4.4 Discriminant validity

The results show that the discriminant validity of the used scales was not very high. Since appeared that the results were threatened by a very high amount of multi collinearity, we investigated the results on the psychometric level as well. The results of the principal component analysis suggested that the constructs of beauty and hedonic quality measured the same underlying latent variable. These results were confirmed by two confirmatory factor analyses. Comparison of the akaike information criterion of a model with five factors and a model with four factors showed that the four model factors model should be preferred. Investigation of the model of fit of the model showed that the model fit of the four factor model is moderate which suggests that the discriminant validity of the rest of the constructs is also only moderate.

We can conclude that results of studies that investigated the relationship between usability and aesthetics, using (among others) the constructs of beauty and hedonic quality are not valid. Having a look at the usability-beauty literature at this moment some studies can be found using these two constructs as separate constructs in their research design. One study that can be named here as an example is an experimental study by Tucha et al. (2012). 80 participants had to use one of four different versions of the same on-line shop, differing in interface-aesthetics (low/high) and interface-usability (low/high). The participants had to find specific items and to rate the shop before and after usage. They had to rate the shop on perceived usability and perceived aesthetics. Among others, participants had to rate the hedonic quality and beauty. Their results show that the interface aesthetics had a medium effect on beauty, in the pre-use phase as well as in the post-use phase. Furthermore, the interface aesthetics had a medium effect on hedonic quality in the post-use phase. However, these results might not be valid as the two constructs measure practically the same.

It is important to realize that there are other HCI research fields which might face low discriminant validity as well. A research field that can be named in this context is research that makes use of the TAM, as introduced in Section 1.2.5. As the scales "perceived usefulness" and "perceived ease-of-use" are quite similar, there is a certain chance that these constructs measure the same latent variable. It is important to evaluate the discriminant validity of these scales in an experimental research setup.

4.5 Conclusions

The good news is that our results showed that the influence of a possible ARS is so small that it can be neglected. This means that HCI researchers do not have to worry about the influence of a possible ARS to their research results. As there are many studies based on correlations of two Likert sclaes, this result is actually a relief. When we would have proved that the ARS indeed has an effect on the research results, this would have meant that a sufficient number of results in the HCI research field were not valid. This leads to the conclusion that a researcher does not have to worry about the influence of the ARS in this research setup. Nonetheless, it was shown that the mixed effects model approach can correct for the ARS. By implementing a random intercept for the different participants, it is corrected for the influence of this repsonse style. The mixed-effects model is superior to the naive analysis in the sense that it assumes that the observations are dependent and therefore does not ignore the data variance structure as the naive analysis does. This overcomes the overestimation of the degrees of freedom and an inflated Type I error. A disadvantage of the mixed-effects model is that the p value, which is a common known and used measure to most HCI researchers, cannot be derived easily in a mixed-effects model – due to the difficult estimation of the degrees of freedom of a mixed-effects model. A possible solution here for is to use Bayesian statistics (Baayen et al., 2008). As, however, the use of p values has been criticized in the statistical literature, this can be seen as a motivation for HCI researchers to get acquainted with the concepts and formalism of Bayesian statistics as well. It goes beyond the scope of this thesis to give an outline of the criticism of the classical view (frequentist statistics, null-hypothesis testing by the means of p values) and the advantages of Bayesian statistics, for further reading see van de Schoot (2010), Klugkist, Wesel, and Bullens (2011) or Hoijtink (2012).

The bad news is that there are other pitfalls for the research design and analysis of usability studies. Based on our research findings, we can give some advice concerning research design and methodology of usability studies, in particularly for researches that focus on the relationship between usability and beauty. These advices are as following:

1. As already suggested by Clark (1973), the researcher should consider the material as well as the subject level of the results of a usability study

2. In a usability study, researchers should certainly use repeated measures (e.g. more than one product)

3. For a mediator analysis, a structural equation modeling model should be used instead of multiple linear regressions

4. The constructs beauty and hedonic quality should not be used as separate constructs but as one common factor

The analysis of the replication of the study by Hassenzahl and Monk (2010) showed a huge amount of covariance between the measured constructs. In this study, different approaches were taken to get more insight into its origin:

- Mixed-effects models
- Psychometric level
- Structural equation modeling

Concerning the mixed-effects models, we tried to explain the covariance in terms of the influence of the ARS. However, it could be concluded that the influence of the ARS was so small that it can be neglected in this research setup. Another approach was to investigate the data on the psychometric level. This showed that the scales beauty and hedonic quality are not indistinguishable and measure the same latent variable. The origin of the covariance in the rest of the scales is still unclear. By the means of structural equation, we finally tested one causal model. Alternative models were however not tested. As a consequence, it was not possible to test this model against other models. For example a model with fluency processing (see Section 4.6.1) would be desirable to test against the model as suggested by Hassenzahl and Monk (2010). Our conclusion is that the origin of the covariance is still ambivalent. It is important to come up with possible explanations to be able to test these in further research.

4.6 Possible explanations and further research directions

When comparing different possible models to explain the relationship between the measured constructs, it could be seen that the research results are threatened by multi collinearity. A confirmatory factor analysis confirmed this and showed that the constructs beauty and hedonic quality are indistinguishable and underlie the same latent

variable. What latent variable could that be then? In the following we will explain the concept of "Fluency of processing" and explain how this could possibly be the underlying latent variable of these constructs.

4.6.1 Processing fluency

The processing of any stimulus can be characterized by a variety of parameters. These parameters are non specific to its content, such as speed and accuracy of stimulus processing (e.g. Reber, Wurtz, & Zimmermann, 2004). It has been shown that these parameters tend to lead to a common experience of processing ease, or stated differently, to a "fluency" of processing (e.g. Clore, 1992; Jacoby, Kelley, Brown, & Jasechko, 1989; Whittlesea & Williams, 1998).

The paper "Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?" by Reber, Schwarz, and Winkielman (2004) provides a first step into a new research direction in which the influence of fluency of processing on aesthetics variables is tested. They review variables known to influence aesthetics judgements, like figural goodness, figure-ground contrast, stimulus repetition, symmetry and prototypicality. They trace the effects of these parameters to changes in processing fluency. They conclude their paper with the conclusion that aesthetic pleasure is a function of the perceiver's processing dynamics: "The more fluently the perceiver can process an object, the more positive is his or her aesthetic response" (Reber, Schwarz, & Winkielman, 2004, p.377). Their proposal entails four specific assumptions:

1. Objects differ in the fluency with which they can be processed
2. Processing fluency is hedonically marked and subjectively experienced as positive
3. The affective response that is elicited by processing fluency results into judgements of aesthetic appreciation; unless the perceiver calls the informational value of his experience into question
4. The effect of processing fluency is moderated by the perceiver's expectations and attribution

These assumptions can be translated to the research design of the current study. The websites can be seen as stimuli which should differ in the fluency with which they can be processed. As suggested by Reber, Schwarz, and Winkielman (2004), this results into

judgements of aesthetic appreciation unless the perceiver calls the informational value of his experience into question. As the participants could only see the screen shot of a website, it is unlikely that they call the informational value into question. Therefore, the high processing fluency probably led to a positive evaluation of the websites. This could explain why the factor of the websites does not explain any variance. The processing fluency might be the latent variable of the two constructs beauty and hedonic quality that turned out to be indistinguishable. A website with a high fluency processing elicited high beauty *and* hedonic quality ratings and a website with a low fluency processing elicited low beauty *and* hedonic quality ratings.

In further research, it has to be tested whether processing fluency is indeed a factor underlying all measured constructs. To test this, one can think of different possible research designs. It is for example possible to conduct a replication of this study. In the replication, the same sample of websites and procedure is used. The only difference is that participants have to rate the websites they get as first and second website again, after the last (tenth) website. By comparing participant's ratings on the beauty and hedonic quality scale on the same websites, one can investigate whether processing fluency indeed affected this ratings. As participants have already seen and rated theses websites, their second ratings should be (dependent on the fluency processing of the websites) higher or lower than the first time. If this is the case, it can be concluded that the processing fluency is indeed the underlying latent variable of the constructs beauty and hedonic quality.

Another way to test the influence of processing fluency would be to manipulate the processing fluency of the websites. It could then be tested whether the websites with a low processing fluency indeed elicit lower ratings on the beauty and hedonic quality scale than the websites with a high processing fluency.

A third possible research setup is a study in which participants have to evaluate websites in *retrospective*. For example, instead of providing participants with the screenshot of the websites and the questionnaire at the same time as we did in this research design, participants have to scan all websites first before they are allowed to fill in the questionnaire. As research has shown that retrospective evaluation is more difficult than real-time evaluation, this could be a way to manipulate the fluency of processing. It would be interesting to investigate whether a replication of this study with the same measures but a retrospective evaluation would gain different research results. When the fluency of processing indeed reflects the latent variable of the construct hedonic quality

and beauty, a possible outcome could be the observation of a *recency effect*. The recency effect, a term coined originally by Hermann Ebbinghaus through studies he performed on himself, refers to the higher recall of the last items of a studied list (Deese & Kaufman, 1957; Murdock, 1962). The recency effect was originally used to support the idea of a short term store but Howard, Kahana, and Sederber (2008) found that −although the effect faded− it was still robustly present in delayed recall. Although more research has to be done in this area, research shows that recency effects are likely to occur in product judgement situation as well (Park & Hastak, n.d.; Park, 1995). When in the hypothetical research design a recency effect can be found, this would mean that the last website participants saw would come more easily to their minds. The processing fluency of this website would be higher than for the rest of the websites. As explained in the last section, a higher fluency processing possibly leads to a higher evaluation of a website. Consequently, a higher evaluation of the last website on the beauty and hedonic quality items would suggest that fluency of processing represents their latent variable.

At last, it could also be interesting to investigate whether the research results are different when the current study is replicated with the same measures and procedure but participants is told beforehand (prior to the actual experiment) that a high covariance between the measured scales can be found and that this observation can be explained by the phenomenon of processing fluency.

4.6.2 Use of implicit methods

Another possible explanation concerns the measurement method: Multiple Likert scales. In the following, we will explain why there is the need to use implicit measurement methods instead.

Robinson and Neighbors (2005) argue why researchers should use implicit methods in personality research and assessment rather than explicit methods, like self-reports. We will give a line-out of their argumentation in the following and then show how this can be translated to the current research.

As implicit methods are based on performance, as for example reaction times, explicit methods, as for example trait measures are based on self-report and therefore require the respondent to have a certain degree of self insight. However, a number of critics show the limitations of this approach. First of all, the history of research on introspection showed that self-reports of mental processes cannot be trusted (MacLeod,

1993). Second, self-reports seem to be based to a large extent on social desirability considerations (Robinson & Neighbors, 2005).

This can be translated to the current research. Likert scales were used to evaluate the different websites, an explicit measure (self-report). As outlined by Robinson and Neighbors (2005), explicit methods have limitations. These limitations can have important implications for the validity of the research results. The most important limitation for our concerns is that the influence of a response style bias and the influence of fluency of processing are method specific for the Likert scale.

The results of the current study show that the beauty and hedonic quality scale measured the same underlying latent variable. The authors of the original study, Hassenzahl and Monk (2010) draw the same conclusion: "We argue that expressive aesthetics and hedonic quality are strongly overlapping constructs" (Hassenzahl & Monk, 2010, p. 25). In a review on empirical research on user experience, Bargas-Avila and Hornbaek (2011) argue that this problem probably originates in the parallel development and publishing of these constructs and shows that a future consolidation may be beneficial. We take a more radical approach. The results suggest that the discriminant validity of the scales is so low that users cannot make a good discrimination between them. Therefore, the current practice of using Likert scales to test usability and its predictors should be replaced by the use of implicit measurement methods. For example the use of reaction times. We conclude these advices with the notion that it is important to keep in mind that these are other measures which may have other biases. When using these alternative measures in an usability study, among others, a research goal should be the assessment of the validity of the measures.

4.7 Limitations of this study

The current study showed that the influence of ARS is not a threat in HCI research. Furthermore, it could be shown that the constructs beauty and hedonic quality are indistinguishable. However, there are also limitations of this study.

These limitations are for the most part of methodological nature. As we replicated the study by Hassenzahl and Monk (2010), these limitations can be seen as a criticism on the research design of the original study.

The constructs of beauty and goodness were measured by the means of a single item only. Despite the practical virtues of the use of single-item measures, there are

a number of psychometric reasons that should make us sceptical about them. First of all, there is consensus that multiple-item scales tend to be more reliable. As the Spearman-Brown formula and classical reliability theory suggests, item responses reflect both random measurement error and true score variance. Therefore, by aggregating over multiple items, errors are cancelled out. This makes the multiple-item measure more reliable than a single-item measure (Robins, Hendin, & Trzesniewski, 2001). Second, by using a multiple-item scale, the content validity for multifaceted constructs is ensured (Robins et al., 2001). The information gained by a single item scale measurement is not sufficient to conduct a confirmatory factor analysis. This results in underestimation. As this was the case in the current study, this point of criticism will be explained elaborately in the following. A requirement for the estimation of a model's parameter is that the model must be identified. A model is seen as "identified" when the information on the sample equals (or exceeds) the needs defined by the estimation of the model parameters. The information about the sample is defined by the unique variances and covariances in the covariance matrix of the observed measures; this information is used to estimate the free model parameters of the factor model (e.g. the factor loadings, the variances, covariances among the factors and the variances and covariances among the errors). One formal rule of thumb for the assessment of the identification of a model is the so called t rule, which states that the number of freely estimated parameters must be less than or equal to the number of unique variance and covariance among the measured variables (Babyak & Green, 2010). However, even if a model passes this rule, the model may still be underidentified – it is important to understand the problem of underidentification conceptually. When there are less then three variables, the fit of the combination of the items stays always the same, there is no "unique" solution. Consequentially, the estimation of the parameters of interest (e.g. factor loadings) is not precise. To be able to calculate the best solution on basis of the data (the best fit), at least three items are necessary. Broadly speaking, this can be understood as a simple problem of algebra. When there are three variables (items) in the calculation, a unique combination of items is possible; therefore the best fit can be estimated.

The aforementioned points of criticism suggests that the single-item measurement of the constructs beauty and goodness is not as reliable and valid as a multiple-item measure would have been. This could have lead to a bigger measurement error and therefore less accurate predictions. Besides, the correlation of two single-item measures generally result in underestimation of the covariance (Neale & Cardon, 2010). The

correlation estimates as gained in Section 3.1 might be biased in the sense that they are yet higher than we found. The same logic applies to the parameter estimate of the regression between beauty and goodness as conducted for the mediator analysis.

It has to be noted here that there is a small amount of researches that suggest that although in theory a multiple-item scale approach should be superior to single-items, in practice there is no difference in the predictive validity of multiple-item and single-item measures (Drolet & Morrison, 2001; Bergkvist & Rossiter, 2007). There are not many researches that explore this direction yet and research results are quite specific as one particular scale was tested (in the first case global self esteem and in the second case attitude toward an advertisements and attitude towards a brand). We cannot be sure if these results can be generalized to scales as used in HCI research. More research has to be done here.

The current study showed that the constructs of beauty and hedonic quality are indistinguishable and underlie the same latent variable. This is another point of criticism. Although we could see that the two constructs measure basically the same, Hassenzahl and Monk (2010) used these two constructs as separate constructs in their research.

Our last point of criticism concerns the sample of the website. In the current study, 10 websites were selected randomly from the pool of websites used in the original study (Hassenzahl & Monk, 2010). All websites have the primary aim of facilitating and supporting sales and on-line transactions to simulate real stores and travel agents. The results from the current study suggests that these websites were so similar that participants could not distinguish between them and ratings were quite similar for all websites. This suggests that the sample of websites was not chosen wisely. For a possible replication of this study, it is advised to use a more broad sample, consisting of websites with different aims. Another approach could be to manipulate the usability of the websites to get a more diverse sample.

References

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Babyak, M., & Green, S. (2010). Confirmatory factor analysis: An introduction for psychosomatic medicine researchers. *Psychosomatic Medicine*, *72*, 587–597.

Bachman, J., & O'Malley, P. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, *48*, 491–509.

Bargas-Avila, J., & Hornbaek, K. (2011). Old wine in new bottels or novel challenges? a critical analysis of empricial studies of user experience. In *Conference on human factors in computing systems.* Vancouver, Canada.

Bates, D. (2005). Fitting linear mixed models in r. *R News*, *5*, 27–30.

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using s4 classes (r package) (.98875-6 ed.) [Computer software manual]. Retrieved from http://cran.r-project.org/web/packages/lme4/index.html

Baumgartner, H., & Steenkamp, J. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*, 143–156.

Baumgartner, H., & Steenkamp, J. (2005). Response biases in marketing research. In R. Grover & M. Vriens (Eds.), *The handbook of marketing research: Uses, misuses and future.* CA: SAGE Publications Inc.

Bergkvist, L., & Rossiter, J. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*(2), 175–184.

Beuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. *Quality & Quantity*, *44*, 761–775.

Billiet, J., & McClendon, M. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*, 608–628.

Bolt, D., & Newton, J. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*, 814–833.

B.Weijters. (2006). *Response styles in consumer research.* Unpublished doctoral dissertation, Ghent University, Ghent, Belgium.

Carifio, J., & Perla, R. (2007). Ten common misunderstandings , misconceptions , persistent myths and urban legends about likert scales and likert response formats

and their antidotes. *Journal of Social Sciences*, *3*(3), 106–116.

Carroll, J. (2009). Human computer interaction (hci). In M. Soegaard & R. Dam (Eds.), *Encyclopedia of human-computer interaction.* Aarhus, Denmark: The Interaction Design Foundation.

Chawda, B., Craft, B., Cairns, P., Rger, S., & Heesch, D. (2005). Do attractive things work better? an exploration of search tool visualisations. In *In preparation.*

Cheung, G., & Rensvold, R. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equation modeling. *Journal of Cross-Cultural Psychology*, *31*(2), 187–212.

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Clore, G. (1992). Cognitive phenomenology: Feelings and the construction of judgment. In L. Martin & A. Tesser (Eds.), *The construction of social judgments* (p. 133-163). Lawrence Eribaum Associates, Inc.

Cunningham, W., Cunningham, I., & Green, R. (1977). The ipsative process to reduce response set bias. *Public Opninion Qaterly*, *41*, 379–394.

Curran, P., West, S., & Finch, J. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16-29.

Davis, F. (1986). *A technology acceptance model for empirically testing new end-user information systems: theory and results.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Massachusetts.

Davis, F. (1993). User acceptance of information technology: System characteristics, user perceptions and behavioral impacts. *International Journal of Man-Machine Studies*, *38*, 475–487.

Deese, J., & Kaufman, R. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, *54*(3), 180–187.

de Jong, M., Steenkamp, J., Fox, J., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*, 104–115.

de Vellis, R. (2003). *Scale development: Theory and applications.* London, United Kingdom: SAGE Publications Inc.

development core team, R. (2007). R: A language and environment for statistical computing [Computer software manual]. Austria, Vienna. Retrieved from http://www.R-project.org

Dooley, D. (2001). *Social research methods*. Upper Saddle River, New Jersey: Prentice-Hall Inc.

Drolet, A., & Morrison, D. (2001). Do we reall need multiple-item measures in service research? *Journal of Service Research*, *3*(3), 196–204.

D.Watson, & Clark, L. (1994). The panas-x. manual for the positive & negative affective schedule-expanded form [Computer software manual]. Iowa. Retrieved from http://www.psychology.uiowa.edu/faculty/clark/panas-x.pdf

England, G., & Harpaz, I. (1983). Some methodological and analytic considerations in cross-national comparative research. *Journal of International Business Studies*, *14*, 49–59.

Ergonomics of human system interaction (Norm No. ISO 9241-11). (1998). ISO, Geneva, Switzerland.

Flaskerud, J. (1988). Is the likert sclae format culturally biased? *Nursing Research*, *37*(3), 185–186.

Galwey, N. (2006). *Introduction to mixed modelling: Beyond regression and analysis of variance*. New York: John Wiley& Sons.

G.Moors. (2003). Diagnosing response style behavior by means of a latent-class factor approach. socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality & Quantity*, *3*, 277–302.

G.Moors. (2012). The effect of response style bias on the measurement of transformational, transaction, and laissez-faire leadership. *European Journal of Work & Organizational Psychology*, *21*, 271–298.

Greenleaf, E. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, *29*, 176–188.

Guilford, J. (1954). *Psychometric methods*. New York: McGraw-Hill.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). Attrakdif: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualitaet [attrakdif: A questionnaire to measure perceived hedonic and pragmatic quality]. In J. Ziegler & G. Szwillus (Eds.), *Mensch & computer 2003. interaktion in bewegung*. Germany, Stuttgart, Leipzig: B.G. Teubner.

Hassenzahl, M., & Monk, A. (2010). The inference of perceived usability from beauty. *Human-Computer Interaction*, *25*(3), 235–260.

Helander, M. (1988). *Handbook of human-computer interaction.* Amsterdam, Holland: Elsevier Science Publishers B.V.

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for the behavioral and social scientists.* New York: CRC-press.

Hornbaek, K. (2006). Current practice in measuring ability: Challenges to usability studies and research. *International Journal of Human Computer Studies*, *64*, 79–102.

Howard, M., Kahana, M., & Sederber, P. (2008). A context-based theory of recency and continguity in free recall. *Psychological Review*, *115*(4), 893–912.

Hox, J., Leeuw, E. D., & Kreft, I. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetzl, & S. Sudman (Eds.), *Measurement errors in surveys.* New York: Wiley.

Hsi-Peng, L., Huei-Ju, Y., & Simon, S. (2001). The effects of cognitive style and model type on dss acceptance: An empirical study. *European Journal of Operational Research*, *131*(3), 649–663.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351-362.

Hui, C., & Triandis, H. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultral Psychology*, *16*, 131–152.

Iabucci, D., Saldanha, N., & Deng, X. (2007). A meditation on mediation: Evidence that structural equations models perform better than regressions. *Journal of Consumer Psychology*, *17*(2), 139–153.

Iacobucci, D. (2009). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, *20*, 90–98.

Jacoby, L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, *56*, 326–338.

Johnson, T., Kulesa, P., Cho, Y., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*, 264–277.

Johnson, T., Kulsea, P., Cho, Y., & Shavitt, S. (2005). The relation between culture and response. *Journal of Cross-Cultural Psychology*, *36*(2), 264–277.

Kaplan, R., & Saccuzzo, D. (2009). *Psychological testing: Principles, applications and issues.* Belmont, CA: Wadsworth.

Kenny, D. A., Kaniskan, B., & McCoach, D. (2011). *The performance of rmsea in models with small degrees of freedom.* (Unpublished)

Kieruj, N., & Moors, G. (2012). Response style behavior: Question format dependent or personal style? *Quality & Quantity*, advance online pubilcation.

Kline, R. (2010). *Principles and practice of structural equation modeling.* London, England: Taylor & Francis Ltd.

Klomp, A. (2011). *De invloed van merkervaring op user experience/eerste indruk vs gebruik [the influence of brand experience on user experience/first impression vs use]* [Unpublished masterthesis]. Enschede, Netherlands.

Klugkist, I., Wesel, F. V., & Bullens, J. (2011). Do we know what we test and do we test what we want to know? *International Journal of Behavioral Development*, *35*, 550–560.

Kurosu, M., & Kashimura, K. (1995). Apparent usability vs. inherent usability: Experimental analysis on the determinants of the apparent usability. In *Conference companion on human factors in computing systems* (pp. 292–293). ACM.

Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models.* Irwin: McGraw Hill.

Lee, C., & Green, R. (1991). Cross-cultural examination of the fishbein behavioral intentions. *Journal of International Business Studies*, *22*, 289–305.

Littel, R., Stroup, W., & Freund, R. (2002). *Sas for linear models.* New York: Wily & Sons.

MacCallum, R., Browne, M., & Sugawara, H. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.

MacLeod, C. (1993). Cognition in clinical psychology: measures, methods or models? *Behaviour Change*, *10*, 169–195.

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*, 1539–1550.

Moorman, R., & Podaskoff, P. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology*, *65*, 131–149.

Moors, G. (2010). Ranking the ratings: A latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research*, *22*, 22–93.

Moran, T., Card, S., & Newell, A. (1983). *The psychology of human-computer interaction.* Boca Raton, Florida: Crc Pr Inc.

Morris, T., & Pavett, C. (1992). Management style and productivity in two cultures. *Journal of International Business Studies*, *23*, 169–179.

Murdock, B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.

Neale, M., & Cardon, L. (2010). *Methodology for genetic studies of twins and families.* Luxembourg: Springer.

Norman, D. (2002). Emotion and design: Attractive things work better. *Interactions Magazine*, *9*(4), 36–42.

N. Reynolds, A. (2010). Assessing the impact of response styles on cross-cultural service quality evaluation: A simplified approach to eliminating the problem. *Journal of Service Research*, *13*, 230–243.

Osgood, C., & Tannenbaum, P. (1957). *The measurement of meaning.* Urbana, IL: University of Illionis Press.

Park, J. (1995). Memory-based product judgments: Effects of presentation order and retrieval cues. *Advances in Consumer Research*, *22*, 159–164.

Park, J., & Hastak, M. (n.d.). Memory-based product judgments: Effects of involvement at encoding and retrieval. *Journal of Consumer Research*, *21*(3), 534–547.

Peterson, R., & Wilson, W. (1992). Measuring customer satisfaction: Fact and artifact. *Journal of the Academy of Marketing Science*, 61–71.

Porat, T., & Tractinsky, N. (2012). It's a pleasure buying here: The effects of web-store design on consumer's emotions and attitudes. *Human Computer Interaction*, *27*, 235–276.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, *8*(4), 364–382.

Reber, R., Wurtz, R., & Zimmermann, T. (2004). Exploring "fringe" consciousness: The subjective experience of perceptual fluency and its objective bases. *Consciousness and Cognition*, *13*, 47–60.

Riordan, C., & Vandenberg, R. (1994). A central question in corss-cultural research: Do employees of different cultures interpret work-related measures in a equivalent manner? *Journal of Management*, *20*, 643–671.

Robins, R., Hendin, H., & Trzesniewski, K. (2001). Measuring global self esteem: Construct validation of a single-item measure and the rosenberg self-esteem scale. *Personality and Social Psychology Bulletin*, *27*(2), 151–161.

Robinson, M., & Neighbors, C. (2005). Catching the mind in action: Implicit methods in personality research and assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology.* Washington: Amer Psychological Assn.

Rorer, L. (1965). The great response-style myth. *Psychological Bulletin*, *63*(3), 129–156.

Saris, W., & Aalberts, C. (2003). Different explanations for correlated disturbance terms in mtmm studies. *Structural Equation Modeling*, *10*, 193–213.

Saris, W., Satorra, A., & Coenders, G. (2004). A new approach to evaluating the quality of measurement instruments. *Sociological Methodology*, *34*, 311–347.

Schaninger, C., & Buss, W. (1986). Removing response-style effects in attributedeterminance ratings to identify market segments. *Journal of Business Research*, *14*, 237–252.

Schmelleh-Engel, K., Moosbrugger, H., & Mueller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

S.Jamieson. (2004). Likert scales: How to ab(use) them. *Medical education*, *38*(12), 1217–1218.

Sokolowski, J., & Banks, C. (2010). *An insightful presentation of the key concepts, paradigms, and applications of modeling and simulation.* Edingburgh, Scotland: John Wiley and Sons.

Tabachnick, B., & Fidell, L. (2005). *Using multivariate statistics.* London: Pearson.

Tractinsky, N. (1997). Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 115–122). ACM.

Triandis, H. (1994). Cross-cultural industrial and organizational psychology. In H. Triandis (Ed.), *Handbook of industrial and organizational psychology.* CA: Consulting

Psychologist Press.

Tucha, A., Rotha, S., Hornbaek, K., Opwisa, K., & Bargas-Avilaa, J. (2012). Is beautiful really usable? toward understanding the relation between usability, aesthetics and effects in hci. *Computers in Human Behavior*, *28*(5), 1596–1607.

van de Schoot, R. (2010). *Informative hypothese: How to move beyond classical null hypothesis testing.* Unpublished doctoral dissertation, University of Utrecht, Utrecht, The Netherlands.

van Herk, H., Poortinga, Y., & Verhallen, T. (2004). Response styles in rating scales: Evidence of method bias in data from six eu countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346–360.

van Rosmalen, J., van Herk, H., & Groenen, P. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, *47*, 157–172.

van Vaerenbergh, Y., & Thomas, T. (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, in press.

Weijters, B., Geuen, M., & Schillewaert, N. (2010). The individual consistency of acquiescence and extreme response style in self report questionnaires. *Applied Psychological Measurement*, *34*(2), 105–121.

Weijters, B., Geuens, M., & Schillewaert, N. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*, 409–422.

Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing*, *26*, 2–12.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods*, *15*(1), 96–110.

Welkenhuysen-Gybels, J., Billiet, J., & Cambr, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of likert-type score items. *Journal of Cross-Cultural Psychology*, *34*, 702–722.

West, S., Finch, J., & Curran, P. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Whittlesea, B. W. A., & Williams, L. (1998). Why do strangers feel familiar, but friends don't? the unexpected basis of feelings of familiarity. *Acta Psychologica*, *98*, 141–166.

Wong, N., Rindfleisch, A., & Burroughs, J. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? the case of the material values scale. *Journal of Consumer Research*, *30*, 72–91.

Appendix A

Mixed effects modeling: Concepts and formalism

The concepts involved in a linear mixed effects model will be introduced by tracing the data analysis path of a simple example.

Imagine the following HCI research example: A researcher wants to investigate whether the perceived usability of a set of websites depends on the perceived creativeness of these websites. Assume an example data set with four participants who had to specify their level of agreement or disagreement for four different websites on six Likert-scale items of which three measure perceived usability and three measure perceived creativeness.

By using R (development core team, 2007), data for this hypothetical research example was simulated. You can see an overview of this hypothetical data set in Table A.1. In the following, it is referred to the participants as $s1$, $s2$ and $s3$ and to the websites as $w1$, $w2$, $w3$ and $w4$. The R code for the simulation and analysis of the data can be found at the end of this section.

Table A.1 is divided into seven sections. In the leftmost sections, you can see subjects, websites and the score for each combination of subject, website and construct (usability and creativeness). The following section lists the fixed effect: the intercept, which is the same for all observations. The right section of the table shows the random effect in this model: The by-subject adjustments to the effect of the scores on the Likert-scale items. For instance, for the first subject, the effect is attenuated by 1 point. The final column lists the residuals, the by-observation noise for each combination.

TABLE A.1: Example of hypothetical data with two variables (usability and creativeness rated for four different websites and random subject intercepts

Subject	Website	Usability	Creativeness	Fixed	Random	Res.
				Int	SubInt	
s1	w1	2.9	2.3	1.8	1.0	-0.3
s1	w2	2.8	2.5	1.8	1.0	-0.5
s1	w3	4.4	2.3	1.8	1.0	1.2
s1	w4	4.1	3.0	1.8	1.0	0.3
s2	w1	3.9	2.9	1.8	-0.3	1.5
s2	w2	0.2	2.8	1.8	-0.3	-2.1
s2	w3	3.9	4.8	1.8	-0.3	-0.3
s2	w4	2.8	3.7	1.8	-0.3	-0.3
s3	w1	5.0	4.0	1.8	0.9	0.5
s3	w2	3.9	3.0	1.8	0.9	0.2
s3	w3	3.6	3.0	1.8	0.9	0
s3	w4	2.5	2.1	1.8	0.9	-0.3

The first four columns are normally available to the researcher (Subjects, websites, scores on the constructs). The remaining columns show the fixed an random effects. Int = Fixed intercept (the same for every participant); SubInt = By-subject adjustments to the intercept (varies for every participant), Res = Residuals.

Formally, this dataset can be summarized as in Equation A.1.

$$\boldsymbol{\gamma}_{ij} = \boldsymbol{\chi}_{ij}\boldsymbol{\beta}_{ij} + \mathbf{S}_{ij}\mathbf{s}_i + \boldsymbol{\varepsilon}_{ij} \tag{A.1}$$

The vector $\boldsymbol{\gamma}_{i_j}$ represents the responses of subject i to item j. In the present hypothetical data set, each of the vectors $\boldsymbol{\gamma}_{i_j}$ comprises scores on the Likert-scale items that measure the usability of the websites. In Equation A.1, $\boldsymbol{\chi}_{i_j}$ is the design matrix, which consists of an initial column of ones and is followed by columns which represent factor contrasts and covariates. As usual for most HCI researches, in this hypothetical research there are no experimental conditions. Therefore, the design matrix of each possible subject-item combination has the form of an identity matrix:

$$\boldsymbol{\chi}_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{A.2}$$

This matrix is the same for all subjects i and items j. It has to be multiplied by the vector of regression coefficients $\boldsymbol{\beta}$. For the present example, this vector takes the form

$$\boldsymbol{\beta} = \begin{pmatrix} 1.8 \\ 0.5 \end{pmatrix} \tag{A.3}$$

in which 1.8 is the coefficient for the intercept and 0.50 is the regression coefficient of the variable creativeness. As the design matrix in this case is an identity matrix, the matrix which results from the multiplication of design matrix and regression coefficients is the same as the regression coefficients matrix,

$$\boldsymbol{\chi}_{ij}\boldsymbol{\beta} = \begin{pmatrix} 1.8 \\ 0.5 \end{pmatrix} \tag{A.4}$$

which shows that the design matrix has no influence. Therefore, we can simplify the suggested model in Equation A.1 as following:

$$\boldsymbol{\gamma}_{ij} = \boldsymbol{\beta}_{ij} + \mathbf{S}_{ij}\mathbf{s}_i + \boldsymbol{\varepsilon}_{ij} \tag{A.5}$$

The purpose of the term $\mathbf{S}_i\mathbf{s}_i$ in Equation A.5 is to make the predictions of the model more precise for the subjects actually examined in the experiment. It represents the random effect of the subjects.

To calculate the $\mathbf{S}_i$ matrix, the design matrix $\boldsymbol{\chi}_{ij}$ is multiplied with a vector specifying for subject i the adjustments that are required for this subject to the intercept. In this case, only the subject intercept is random. As there are no conditions involved, the second value of the vector stays 0.9 for each combination of subject and items. The second value of the vector would be different for every participant, when conditions would be involved in the experimental design of the experiment. However, this case is rare in HCI research and therefore not considered here.

For the last subject in Table 1 this would be:

$$\mathbf{S}_{3j}\ \mathbf{s}_3 = \begin{pmatrix} 0.89 \\ 0.9 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2.9 \\ 0.9 \end{pmatrix} \tag{A.6}$$

When we take a closer look at this vector, we can conclude that the intercept for the last subject has to be adjusted upwards by 0.89. This suggests that this respondent scored generally quite low on the Likert-scale items.

Note that in this example there is only one random effect: the subject random effect. Other models can be thought of where one wants to bring also the item random

effect into the model. For readers who are interested in the application of an item random effect, the article by Baayen et al. (2008) is advised for further reading.

The last term of Equation A.5 specifies the vector of the residual errors $\boldsymbol{\epsilon}_{ij}$.

When one substitutes all values in to Equation A.5, this results in the following matrix for participant 3:

$$\boldsymbol{\gamma}_3 = \boldsymbol{\gamma}_{3j} = \boldsymbol{\beta}_{3j} + \mathbf{S}\mathbf{s}_3 + \boldsymbol{\epsilon}_{1j} = \begin{pmatrix} (1.8 + 0.50) + (0.89 + 0.9) + 0.5 \\ (1.8 + 0.50) + (0.89 + 0.9) + 0.2 \\ (1.8 + 0.50) + (0.89 + 0.9) + 0 \\ (1.8 + 0.50) + (0.89 + 0.9) + (-0.3) \end{pmatrix} \tag{A.7}$$

Written in matrix from, this leads to the general model specification

$$\boldsymbol{\gamma} = \boldsymbol{\beta} + \mathbf{W} + \boldsymbol{\epsilon} \tag{A.8}$$

To complete all model specifications, we have to be precise about the structure of the random effects of our data set. A random variable is defined as a normal variate with zero mean and unknown standard deviation. The estimates for the standard deviations of the four random effects for this hypothetical data set are $\hat{\sigma}_{s(int)} = 0.84$ for the by-subject adjustments to the intercept and $\hat{\sigma}_{\epsilon} = 1.0$ for the residual error.

With these four random effect parameters we can complete the model specification and present the full formal specification of the corresponding mixed-effects model:

$$\boldsymbol{\gamma} = \boldsymbol{\beta} + \mathbf{Z} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \sigma^2), \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\Sigma), \mathbf{b} \perp \boldsymbol{\epsilon}, \tag{A.9}$$

where $\sim$ represents the relative variance-covariance matrix for the random effects. The symbol $\perp$means that the random variables are independent and N indicates the multivariate normal (Gaussian) distribution.

A.1 R code

```
#Simulation of data to give an example in the introduction section
library(lme4)

#I. Simulation of data
#5p Likert scale, participants have rated usability for each website (=4)
usability_s1 <- as.vector(rnorm(4, mean = 4.5, sd = 1)) #subject 1
usability_s2 <- as.vector(rnorm(4, mean = 3.5, sd = 1)) #subject 2
usability_s3 <- as.vector(rnorm(4, mean = 4, sd =1)) #..]

#I forgot to set the seed! Therefore, I saved the simulated data here.
usability <- as.matrix(c(2.9,2.8, 4.4, 4.1, 3.9, 0.2,
                         3.9, 2.8, 5.0, 3.9, 3.6, 2.5))

#create an effect to get an impression of how the data must look like
#(to get an RST effect, there must be an effect in the data. I chose an
#effect of 0.3 to add radom error. Rnorm adds random noise.
#Simulate data
creativity <- 0.3*usability + rnorm(12, 2, 1)

#Again, I save my values (as the seed was not set)
creativity2 <- as.matrix(c(2.3, 2.5, 2.3, 3.0, 2.9, 2.8,
                           4.8, 3.7, 4.0, 3.0, 3.0, 2.1))

#All variables
subjects <- as.matrix(c(rep(1,4), rep(2,4), rep(3,4)))
creativity2;usability

#II. The design matrix (in this case its an identity matrix, as
#there is only one condition!)
xij <- matrix(c(1, 0, 0, 1), nrow = 2, ncol = 2)

#II. Beta, regression coefficient
#1)Via matrix algebra
x <- creativity2
intercept <- as.matrix(cbind(1, x)) #add intercept+make matrix
beta <- (solve(t(intercept)%*%intercept))%*%(t(intercept)%*%usability)

#Check with in-built function
model <- lm(usability~creativity2)
coef(model) #gives the same values
summary(model)

#III. Multiplicate design matrix with coef (stays the same of course)
coef <- as.matrix(c(1.8,0.5))
xijbeta <- xij%*%coef

```

```
47 #Analysis
48 model2<-lmer(usability~creativity2 +  (1|subjects))
49 summary(model2)
50 coef(model2)
51 residuals(model2)
```

Appendix B

Data simulation (R code)

```
###Graphic for introduction
library(MASS)
set.seed(100)

#Simulate data, r = 0.7, mean is 4
xy <- mvrnorm(20, mu = c(4,5),
              Sigma = matrix(c(1,0.7, 0.7,1),2,2))
cor(xy[,1],xy[,2]) # r = 0.7

#Simulate data with a smaller correlation (r = 0.4),
xy2 <- mvrnorm(20, mu =c(4,5),
               Sigma = matrix(c(1,0.3,0.3,1),2,2))
cor(xy2[,1], xy2[,2]) #correlatie is 0.3

#Make plot
plot(xy[,1],xy[,2], xlab = "Attractiveness",
     ylab = "Usability", main = "Effect of ARS")
points(xy2[,1], xy2[,2], pch=20, col="red")
abline(lm(xy[,1]~xy[,2]))
abline(lm(xy2[,1]~xy2[,2]), col = "red")
```

Appendix C

Power analysis (R code)

```
#Approximate the power by simulation (of mediator analysis)
############################################################

#Function calculates the power of a bivariate correlation (y,x)
#by simulating data -numsim- times
#& dividing all p values <0.05 by the number of
#simulations.(can be tested one or two sided)

#Input: numsim -> number of simulations,
#var -> bivariate normal distribution with r = x effect
############################################################

power_correlation <- function (numsim,var, n, side = "one-sided")

{                   r <- matrix(0, numsim, 1) #for memory
                    t <- matrix(0, numsim, 1)
                    p <- matrix(0, numsim, 1)
  degrees_of_freedom <- (length(var[,1])+length(var[,2]))-2


   for(i in 1:numsim)

{
            r[i] <- cor(var[,1],var[,2]) #correlation
            t[i] <- r[i]*sqrt((n-2)/1-(r[i]^2)) #t values
if(side=="one-sided") p[i] <- 1-pt(t[i], df = degrees_of_freedom)
if(side=="two-sided") p[i] <- 2*(1-pt(t[i], df = degrees_of_freedom))
  }

              mean_t_values <- round(mean(t), digits = 5) #mean t values
             mean_p_values <- round(mean(p), digits = 5) #mean p values
                     power <- sum(p < 0.05)/numsim #power

```

```
34 par(mfrow=c(1,2))
35 hist(p, main = "P values (numsim simulations)", xlab = "", nclass = 10)
36 abline(v = mean_p_values, col="blue")
37 hist(t, main = "T values (numsim simulations)", xlab = "", nclass = 10)
38 abline(v = mean_t_values, col = "red")
39
40              output <-  list(mean_t_values = mean_t_values,
41                              mean_p_values = mean_p_values,
42                              power = power)
43
44 names(output) <- c("Mean T values","Mean p values", "Power")
45 format(output, trim = FALSE, justify = c("centre"))
46 }
47
48 #Without response style bias (effect 0.3)
49 y = runif(20, min =1, max =7)
50 x = 0.3*y + rnorm(20,3,1)
51 var <- cbind(y,x)
52 power_correlation(1000,var,20)
53
54 #With simulated response style bias(mu=4 in both variables)(effect = 0.3)
55 library(MASS)
56 var <- mvrnorm(20, mu = c(4,4), Sigma = matrix(c(1,0.3, 0.3,1),2,2))
57 var2 <- mvrnorm(200, mu =c(4,4), Sigma = matrix(c(1,0.3,0.3,1),2,2))
58 power_correlation(100,var,20)
59 power_correlation(100,var2, 200)
60
61
62
63 #Simulation for power regression (in article it's a mediator analysis,so
64 #more than one regression calculations are involved, so this is only an
65 #approach to get at least some idea about the sample size I think that a
66 #simulation study for a whole mediator is quite
67 #time consuming to make..I used the in built function
68 #(see beneath), but I am not sure about the beta and sigma values
69 #(that's the reason why I also made my own function,
70 #because then I know exactly what is going on and have at least some
71 #idea about the sample size..)
72
73 power_f_test <- function (numsim, y, x)
74 {
75                         p <- matrix(0, numsim,1)    #for memory
76                         f <- matrix(0, numsim, 1)
77                     power <- matrix(0, numsim, 1)
78                     adj.R <- matrix(0, numsim, 1)
79
80  for(i in 1:numsim)
81 {
```

```
82            int <- as.matrix(cbind(1, x))
83
84           coef <- solve(t(int)%*%int)%*%(t(int)%*%y) #regression coefficients
85            fit <- int%*%coef #fitted values
86            MSS <- var(fit) #mean sum of squares
87            TSS <- var(y) #total sum of squares
88            RSS <- TSS-MSS #residual sum of squares
89              R <- MSS/TSS #R
90       adj.R[i] <- 1-(((length(y)-1)/(length(y)-ncol(int)))*(1-R))#adjusted R
91           f[i] <- (MSS/ncol(x))/(RSS/(length(y)-ncol(int)))#F value
92           p[i] <- pf(f[i], ncol(x), (length(y)-ncol(int)), lower.tail = FALSE)#p
93  }
94
95       mean_f_values <- round(mean(f), digits = 5) #mean f values
96       mean_p_values <- round(mean(p), digits = 5) #mean p values
97          mean_adj.R <- round(mean(adj.R), digits = 5) #mean adj. R values
98               power <- sum(p < 0.05)/numsim #power
99
100 par(mfrow=c(1,2))
101 hist(f, main = "F values (numsim simulations)",
102      xlab = "", nclass = 10)
103 abline(v = mean_f_values, col="red")
104 hist(p, main = "P values (numsim simulations)",
105      xlab = "", nclass = 10)
106 abline(v = mean_p_values, col="blue")
107
108              output <-  list(mean_f_values = mean_f_values,
109                              mean_p_values = mean_p_values,
110                              mean_adj.R = mean_adj.R,
111                              power = power)
112
113 names(output) <- c("Mean F values","Mean P values",
114                    "Mean Adjusted R ","Power")
115 format(output, trim = FALSE, justify = c("centre"))
116 }
117
118 #Without response style bias (effect 0.3)
119 y <- as.matrix(runif(20, min =1, max =7))
120 x <- as.matrix(0.3*y + rnorm(20,3,1))
121
122 power_f_test(1000,y,x)
123
124 #For the whole mediation, I used the package powerMediation. H
125 #owever I am not sure about the beta and sigma
126 #values.. (that's the reason why I wanted to make my own simulation function)
127 library(powerMediation)
128 powerMediation.VSMc(200, 4.5, sigma.m = 1, sigma.e = 1, 0.3,
129                     alpha = 0.05,verbose = TRUE)
```

Appendix D

Items

Table D.1: Hedonic quality (HQ)

Original (German) Hassenzahl et al. (2003)	English translation Hassenzahl and Monk (2010)	Dutch translation Klomp (2011)
Stillos-Stilvoll	Tacky-Stylish	Stijlloos-Stijlvol
Minderwertig-Wertvoll	Cheap-Premium	Minderwaardig-Waardevol
Phantasielos-Kreativ	Unimaginative-Creative	Fantasieloos-Creatief
Lahm-Fesselnd	Dull-Captivating	Saai-Fascinerend

Table D.2: Pragmatic quality (PQ)

Original (German) Hassenzahl et al. (2003)	English translation Hassenzahl and Monk (2010)	Dutch translation Klomp (2011)
Kompliziert-Einfach	Complicated-Simple	Ingewikkeld-Eenvoudig
Unpraktisch-Praktisch	Impractical-Practical	Onpraktisch-Praktisch
Unberechenbar-Voraussagbar	Unpredictable-Predictable	Onvoorspelbaar-Voorspelbaar
Verwirrend-Uebersichtlich	Confusing-Clearly structured	Verwarrend-Overzichtelijk

Table D.3: Goodness and beauty

Original (German) Hassenzahl et al. (2003)	English translation Hassenzahl and Monk (2010)	Dutch translation Klomp (2011)
Haesslich-Schoen	Ugly-attractive	Lelijk-Mooi
Schlecht-Gut	Bad-Good	Slecht-Goed

Table D.4: Independent scale

English version	German translation	Dutch translation
Moonily-Alert	Vertraeumt-Wachsam	Dromerig-Alert
Inattentive-Attentive	Unaufmerksam-Aufmerksam	Onoplettend-Oplettend
Distracted-Concentrating	Abgelenkt-Konzentriert	Afgeleid-Geconcentreerd
Undetermined-Determined	Unentschlossen-Entschlossen	Onbeslist-Beslist

Appendix E

Questionnaire

E.1 First screen

Welkom bij dit onderzoek naar de factoren die invloed hebben op de usability van websites. Je krijgt zometeen de screenshot van een website te zien. De bedoeling is dat je deze kort bekijkt en dan aan de hand van je eerste impressie een korte vragenlijst invult.In totaal zul je 10 websites beoordelen. Voordat je de eerste screenshot kan zien, zul je eerst nog om een aantal gegevens worden gevraagd (leeftijd,etc.) en nadat je alle websites hebt beoordeeld, zul je nog een aantal vragen over jezelf beantwoorden.
Aan het begin van de vragenlijst kan je de taal instellen (zie drop down menu aan de bovenkant van de website). Om resultaten te krijgen die zo valide mogelijk zijn, wil ik je vragen om hier je eigen moedertaal te kiezen (ook al ben je Duits en spreek je vloeiend Nederlands!).
Druk op de button om naar de vragenlijst te komen.
Bedankt voor je deelname!

E.2 Screenshot of first questions

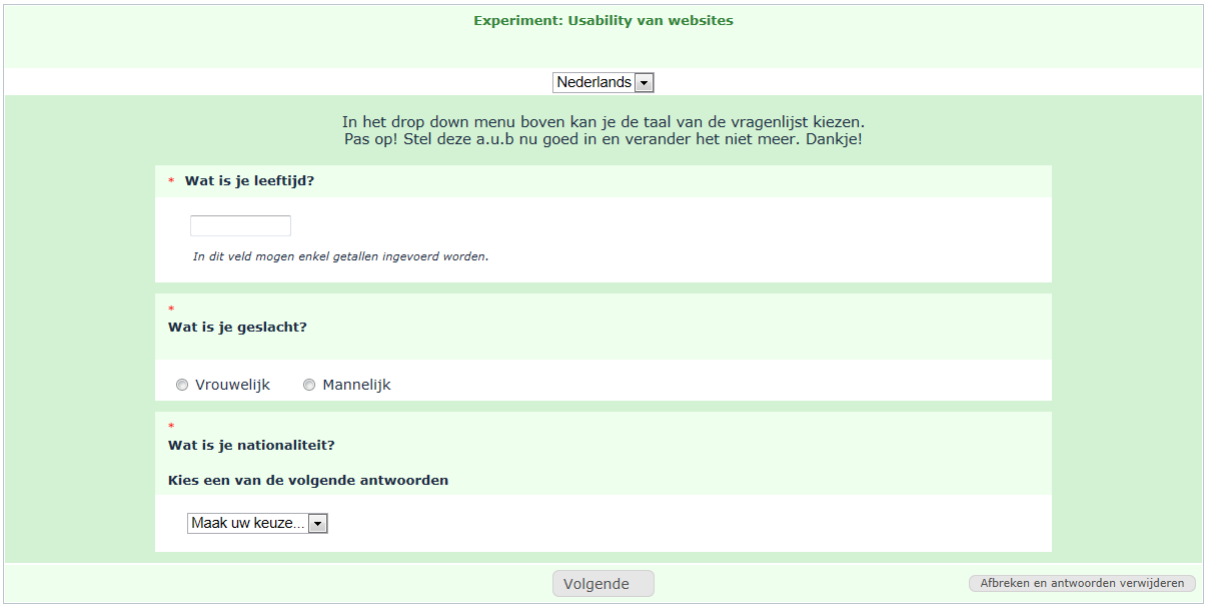

FIGURE E.1: First questions

E.3 Screen shot of an example site

FIGURE E.2: Example site

E.4 Screen shot of the independent scale

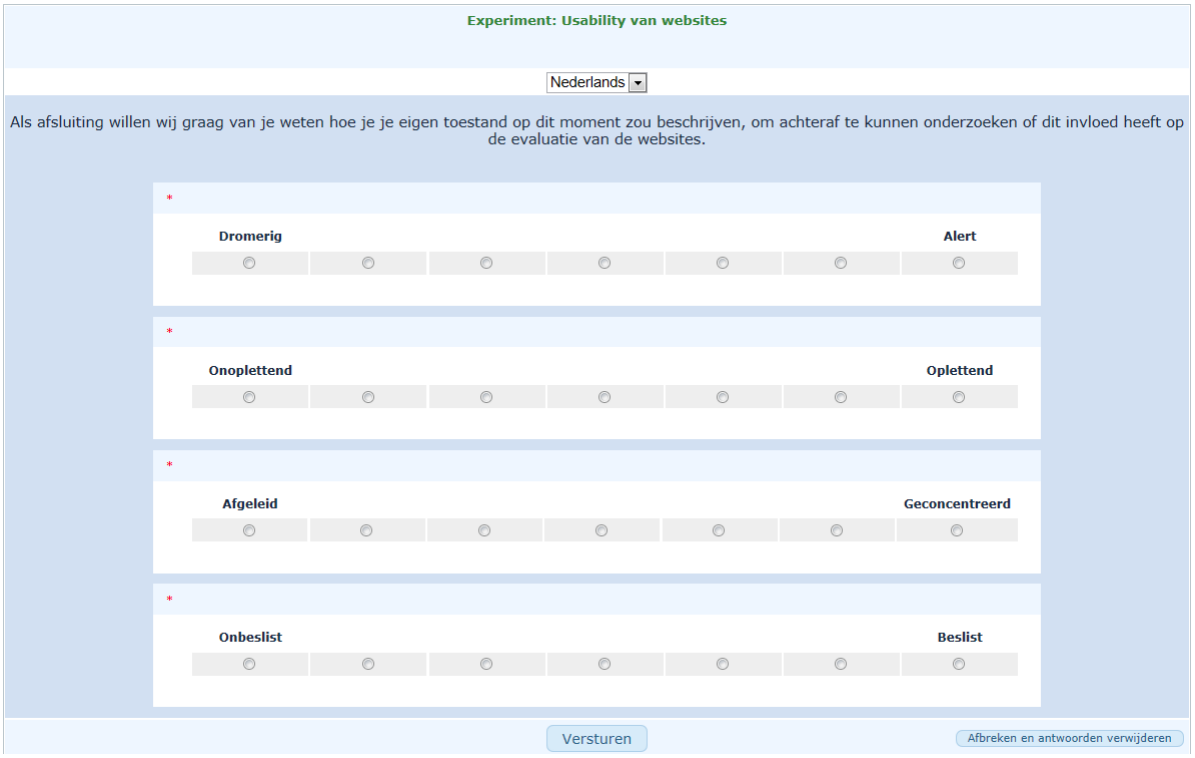

FIGURE E.3: Independent scale

Appendix F

Randomization of the websites (Python code)

```
1  #! /usr/bin/python
2  print "Content-Type: text/html\n\n"
3
4
5  import MySQLdb
6  import cgitb; cgitb.enable();
7  import random
8
9  from pprint import pprint
10 conn = MySQLdb.connect (host = "svn.blinkt.de",
11                           user = "ingazitrone",
12                         passwd="*", db = "ingazitrone")
13
14 cursor = conn.cursor ()
15
16 def wuerfel(lang):
17     cursor.execute ("select gid, group_name from lime_groups where sid='465776'\
18                     and language=%s", lang)
19     foo=  list(cursor.fetchall())
20
21     first = None
22     last = None
23     for i in foo:
24         if i[1] == "Erste Fragen" or i[1] == "Eerste vragen":
25             first = i
26         if i[1] == "Laatste vragen" or i[1] == "Letzte Fragen":
27             last = i
28
29     foo.remove(first)
30     foo.remove(last)
```

```
31      random.shuffle(foo)
32
33
34      alles = (first,) + tuple(foo) + (last,)
35
36      print "<ol>\n"
37      for i,(gid, name) in enumerate(alles):
38          print "<li>%d: %s</li>" % (gid,name)
39          cursor.execute ("update lime_groups set group_order=%d where gid=%d"
40                          %( i, gid))
41
42      print "</ol>"
43      return alles
44
45
46  def machgleich(alles):
47      cursor.execute ("select gid, group_name from lime_groups where sid='465776'\
48                      and language=%s","de");
49      foo=  list(cursor.fetchall())
50      pos = {}
51
52      for i in foo:
53          if i[1] == "Erste Fragen" or i[1] == "Eerste vragen":
54             pos[i] = 0
55          if i[1] == "Laatste vragen" or i[1] == "Letzte Fragen":
56              pos[i] = len(foo)
57
58          for jidx,j in enumerate(alles):
59              if i[1] == j[1]:
60                  pos[i] = jidx
61      print "<ul>\n"
62      for (gid,name),idx in pos.items():
63          print "<li>Index %d: %d: %s</li>" % (idx,gid,name)
64          cursor.execute ("update lime_groups set group_order=\
65                          %d where gid=%d" %( idx, gid))
66      print "</ul>"
67
68
69
70
71  alles =wuerfel("nl")
72  machgleich(alles)
73
74
75
76  cursor.close ()
77  conn.close ()
```

Appendix G

Websites

http://www.iwantoneofthose.com

http://www.boden.co.uk

http://www.pashmina-pashminas.co.uk

http://www.beachcombertours.co.uk

http://www.play.com

http://www.paramountzone.com

http://www.eveningdresses.co.uk

http://www.boystoys.com

http://www.countrybookshop.co.uk

http://www.archersdirect.co.uk

Appendix H

Analysis (R code)

H.1 Data preparation

```
### I. Preparing the data for further analysis

#Load packages
library(foreign)
library(reshape2)

#Import data
setwd("C:/Users/Inga/Documents/thesis/data")
HPG <- as.data.frame(read.spss("alldata.sav"))

#Control data/data format
head(HPG)
tail(HPG)
summary(HPG)
head(HPG[9:108])

#Reshaping on website ratings
HPG.melt <- melt(HPG[,c(1,9:108)], id=c("id"))

HPG[,c(1,9:108)]
head(HPG.melt)
tail(HPG.melt)

#Separating website and item
website <- c("arche", "beach", "boden", "boys", "country",
"evening", "iwant", "paramount", "pashima", "play")

quest <- c(paste(c("HQ"),1:4, sep=""),
paste(c("PQ"),1:4, sep=""),"Goodness", "Beauty")

```

```
#Making all combinations
HPG.long <- expand.grid(HPG$id,quest, website)
head(HPG.long)
tail(HPG.long)

#Completing the long data frame
names(HPG.long) <- c("Subj", "Item", "Website")
head(HPG.long)
head(HPG.long$Item)
HPG.long$scale


#Adding the scale (for aggregation)
HPG.long$Scale <- substr(HPG.long$Item,1,1)

#Adding the values, finally
HPG.long$Rating <- HPG.melt$value

## Subject level analysis ####
#Collapsing over websites and items
HPG.subj <- dcast(Subj ~ Scale, mean, data=HPG.long)
#Collapsing over websites (for factor analysis)
HPG.fa <- dcast(Subj ~ Item, mean, data=HPG.long)
HPG.fa.grouped <- dcast(Subj+Website ~ Item,
                        mean, data=HPG.long)

## Adding the independant scales
HPG.subj <- cbind(HPG.subj, HPG[,109:112])
HPG.fa <- cbind(HPG.fa, HPG[,109:112])

## Material level analysis ####
## Collapsing over Subj and items
HPG.material <- dcast(Website ~ Scale, mean,
                      data=HPG.long)

#For mixed effects analysis
HPG.crossed <- dcast(Subj+Website ~ Scale, mean,
                     margins="Rating", value.var="Rating",
                     data=HPG.long)
names(HPG.crossed[4]) <- "Rating"
```

H.2 Data exploration

```
#II. Data exploration

#Load packages
library(ggplot2)
library(lme4)
library(LMERConvenienceFunctions)
library(psych)
library(languageR)
library(micEcon)
library(MuMIn)

source("data_preparation.R")

#Make one scale of independent scale items
dim(HPG.subj)
HPG.subj$I <- apply(HPG.subj[6:9,], 1, mean)


#Exploring Intercorrelations
##Aggregated over subjects
plotmatrix(data=HPG.material[,2:5]) + geom_smooth(method=lm)+
  theme_bw(base_size=20)
cor(HPG.material[,2:5])

##Aggregated over websites (Independent collapsed into one scale)
plotmatrix(data=HPG.subj[,c(2:5,10)]) +geom_smooth(method=lm)+
  theme_bw(base_size=20)
cor(HPG.subj[,c(2:5,10)])

summary(m1 <- lm(P~I, HPG.subj))
summary(m2 <- lm(H~I, HPG.subj))
summary(m3 <- lm(G~I, HPG.subj))
summary(m1)

#The response style bias can be found for all scales as used in
#the article (B/P/G), but not for the independent scale.

#Naive analysis
plotmatrix(data=HPG.crossed[,3:6])
+geom_smooth(method=lm)+theme_bw(base_size=20)
cor(HPG.crossed[,3:6]) #smallest correlation is 0.


#Model comparison
#In the article the relationship between P -> B was tested.
#In the following it is tested which is the best model here for.
```

```
47
48 ##Pragmatism, Website as fixed effect:
49 model_U4 <- lm(P ~ G + Website, data=HPG.crossed)
50 dredge(model_U4)
51 model2 <- lm(P ~ B + H + G + Website, data = HPG.crossed)
52 summary(model2)
53 dredge(model2)
54 #This shows that common factor of the websites does not contribute
55 #to the explained variance.
56
57 #Therefore, we concentrate on P -> B/H/G
58 model0 <- lm(P ~ B, data=HPG.crossed)
59 model1 <- lm(P ~ B + H, data=HPG.crossed)
60 model2 <- lm(P ~ B + H + G, data=HPG.crossed)
61 model3 <- lm(P ~ H + G, data = HPG.crossed)
62 model4 <- lm(P ~ G, data = HPG.crossed)
63
64
65 ModelComparison <- function(y){
66   m1 <- c(summary(lm(y ~ B, data = HPG.crossed))$adj.r.squared,
67           summary(lm(y ~ B, data = HPG.crossed))$fstatistic[1])
68   m2 <- c(summary(lm(y ~ B + H, data = HPG.crossed))$adj.r.squared,
69           summary(lm(y ~ B + H, data = HPG.crossed))$fstatistic[1])
70   m3 <- c(summary(lm(y ~ B + H + G, data = HPG.crossed))$adj.r.squared,
71           summary(lm(y ~ B + H + G, data = HPG.crossed))$fstatistic[1])
72   m4 <- c(summary(lm(y ~ H + G, data = HPG.crossed))$adj.r.squared,
73           summary(lm(y ~ H + G, data = HPG.crossed))$fstatistic[1])
74   m5 <- c(summary(lm(y ~ G, data = HPG.crossed))$adj.r.squared,
75           summary(lm(y ~ G, data = HPG.crossed))$fstatistic[1])
76   output<-list(m1,m2,m4, m5)
77   names(output)<-c("adj. R /F Model1","adj. R /F Model2",
78                    "adj. R /F Model3")
79   format(output, trim = FALSE, justify = c("centre"))
80 }
81
82 ModelComparison(HPG.crossed$P)
83
84 AIC(model0,model1,model2,model3,model4)
85 summary(model3)
86 AIC(model3)
87
88 #The F value & adjusted R square is the biggest for model3
89 #(P ~  H + G), which suggests that this is the best model
90 #this is confirmed when we look at the ACI values
91 #(lowest for this model)
92
93 #Automated model comparison
94 comp <- dredge (lm (P ~ B*H*G, HPG.crossed))
```

```
95  get.models(comp)
96  #Paramter values are instable. This suggests that there is
97  #multicollinearity. This suggestion is tested later on.
98
99  #Mixed effects models
100
101 ## MS: Predicting Pragmatism: ####
102 ##Subject level (random effect is ARS)
103 model_U <- lmer(P~B+H+G+(1|Subj),HPG.crossed)
104 comp2 <- dredge(model_U)
105 get.models(comp2)
106
107 model_U7 <- lmer(P~G+H+(1|Subj),HPG.crossed)
108 model_U3 <- lmer(P~G+(1|Subj),HPG.crossed)
109 anova(model_U3, model_U7)
110 summary(model_U7)
111 ## Moderate standard deviation of ARS random effect
112
113 ##Material level
114 #(random effect is the common factor of websites, CF)
115 model_W <- lmer(P~B+G+H+(1|Website),HPG.crossed)
116 comp3 <- dredge(model_W)
117 get.models(comp3)
118
119 model_W7 <- lmer(P~G+H+(1|Website),HPG.crossed)
120 model_W3 <- lmer(P~G+(1|Website),HPG.crossed)
121 anova(model_W3, model_W7)
122 summary(model_W7)
123 ## Low standard deviation on CF random effect
124
125 ## All possible models
126 modelB <- lmer(B ~ P*G*H +(1|Subj), data=HPG.crossed)
127 dredge(modelB)
128 modelP <- lmer(P ~ B*G*H +(1|Subj), data=HPG.crossed)
129 dredge(modelP)
130 modelH <- lmer(H ~ B*G*P +(1|Subj), data=HPG.crossed)
131 dredge(modelH)
132 modelG <- lmer(G ~ B*P*H +(1|Subj), data=HPG.crossed)
133 dredge(modelG)
134
135 #Compare naive model & mixed effects model
136 mixed <- lmer(P ~ H + G + (1|Subj), HPG.crossed)
137 naive <- lm(P ~ H + G, HPG.crossed)
138
139 #Plot intercepts & resid(naive model)
140 coef <- coef(mixed)$Subj
141 coef2 <- coef[,1]
142 par(mfrow = c(1, 2))
```

```
143 plot(histogram1);plot(histogram2)
144 histogram1 <- hist(resid(mixed))
145 histogram2 <- hist(coef2)
146
147 plot(histogram1, main = "Naive model",
148      xlab = "Residuals")
149 plot(histogram2, main = "Mixed model",
150      xlab = "Random intercepts")
151
152 #When 2nd histogram is centered to 0
153 histogram2 <- hist(coef2, breaks = c(-5,0,5))
154 plot(histogram1, main = "Naive model",
155      xlab = "Residuals")
156 plot(histogram2, main = "Mixed model",
157      xlab = "Random intercepts")
158
159 #Compare paramter estimates
160 #Paramter estimates are almost the same (fixed effects)
161 summary(mixed)
162 summary(naive)
163
164 #Compare predicted values
165 pred_naive <- fitted(naive)#pred.val.naive model
166 pred_mixed <- fitted(mixed)
167
168 #Compare predicted values
169 head(pred_mixed)
170 head(pred_naive)
171 dif <- pred_mixed-pred_naive
172 #You can see that there are differences between the predicted
173 #values of both models.However,
174 #1)these differences are quite small (see dif)
175 #2)the predicted values are sometimes smaller and sometimes
176 #bigger for the mixed effects model and not generally bigger
177 #as would be expected in case of a ARS
178
179 #Compare MSEP of both models
180 MSEP_naive <- mean((HPG.crossed$P-pred_naive)^2)#0.94
181 MSEP_mixed <- mean((HPG.crossed$P-pred_mixed)^2)#0.81
182
183 #Calculate R square
184 rSquared(HPG.crossed$P, resid(naive))#0.38
185 rSquared(HPG.crossed$P, resid(mixed))#0.47
186
187 #(check this result)+caclulate corrected R square
188 #1)for the naive model:
189 MSSnaive <-var(pred_naive) #mean sum of squares
190 TSSnaive <-var(HPG.crossed$P) #total sum of squares
```

```
RSSnaive <-TSSnaive-MSSnaive #residual sum of squares
Rnaive <- MSSnaive/TSSnaive #R square
adjr_naive<- 1-(((length(HPG.crossed$P)-1)/
            (length(HPG.crossed$P)-ncol(int)))*(1-Rnaive))#adjusted R
#2)for the mixed model:
MSSmixed <-var(pred_mixed) #mean sum of squares
TSSmixed <-var(HPG.crossed$P) #total sum of squares
RSSmixed <-TSSmixed-MSSmixed #residual sum of squares
Rmixed <- MSSmixed/TSSmixed #R square
x <- cbind(HPG.crossed$H, HPG.crossed$G + (HPG.crossed$Subj))
int <- as.matrix(cbind(coef2, x))
adjr_mixed <- 1-(((length(HPG.crossed$P)-1)/
  (length(HPG.crossed$P)-ncol(int)))*(1-Rmixed))#adjusted R

#Compare resdiuals
res_naive <- residuals(naive)
res_mixed <- residuals(mixed)#function not available for lmer
head(res_naive)
head(res_mixed)
diffr <- res_mixed - res_naive
#differences seem to be quite random

#Investigate collinearity of the variables
#1)Via function written by
#https://raw.github.com/aufrank/R-hacks/master/mer-utils.R

vif.mer <- function (fit) {
  ## adapted from rms::vif

  v <- vcov(fit)
  nam <- names(fixef(fit))

  ## exclude intercepts
  ns <- sum(1 * (nam == "Intercept" | nam == "(Intercept)"))
  if (ns > 0) {
    v <- v[-(1:ns), -(1:ns), drop = FALSE]
    nam <- nam[-(1:ns)]
  }

  d <- diag(v)^0.5
  v <- diag(solve(v/(d %o% d)))
  names(v) <- nam
  v
}

vif.mer(modelP)
#Value should not be higher than 5.

```

```
#Via function written by
#http://yatani.jp/HCIstats/MultilevelLinear#Multicollinearlity
panel.cor <- function(x, y, digits=3, prefix="", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- cor(x, y,use="complete.obs")
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  prefix <- "r = "
  prefix2 <- "\nCI lower = "
  prefix3 <- "\nCI upper = "
  prefix4 <- "\np = "
  rc <- cor.test(x,y)
  rci <- rc$conf.int
  rcp <- rc$p.value
  star <- symnum(rcp, corr = FALSE, na = FALSE,
                 cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
                 symbols = c("***", "**", "*", ".", " "))
  txt2 <- format(c(rci, 0.123456789), digits=digits)[1]
  txt3 <- format(c(rci, 0.123456789), digits=digits)[2]
  txt4 <- format(c(rcp, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, prefix2, txt2, prefix3, txt3,
               prefix4, txt4, " ", star, sep="")
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = 1)
}

pairs(HPG.crossed[,3:6], lower.panel=panel.smooth,
      upper.panel=panel.cor)

#almost the same graph as:
plotmatrix(data=HPG.crossed[,3:6]) +geom_smooth(method=lm)
+theme_bw(base_size=20)
```

```
library(ggplot2)
library(psych)
library(lavaan)
library(qgraph)
source("data_preparation.R")

## HPG.fa is collapsed over websites, hence any effect is purely
##personal preferences of subjects
fit <- principal(HPG.fa[,2:15], nfactors=3, rotate="varimax")
fit # print results
## Beauty is indistinguishable of HQ
## Goodness more strongly related to HQ/B, but also some link to PQ

fit <- factanal(HPG.fa[,2:15], factors=3)
```

```
15 print(fit, digits=2, cutoff=.3, sort=TRUE)
16
17 # plot factor 1 by factor 2
18 load <- fit$loadings[,1:3]
19 plot(load,type="n") # set up plot
20 text(load,labels=names(HPG.fa[,2:15]),cex=.7) # add variable names
21
22 ## Confirmatory factor Analysis
23 ## a five factors model
24 cfa.model1 <- 'Pragmatism =~ PQ1 + PQ2 + PQ3 + PQ4
25               Hedonism =~ HQ1 + HQ2 + HQ3 + HQ4
26               ASR =~ indep1 + indep2 + indep3 + indep4
27               B =~ Beauty
28               G =~ Goodness'
29
30 cfa1 <- cfa(cfa.model1, data=HPG.fa)
31 summary(cfa1)
32
33
34 ## Covariance table indicates that P is almost independent of t
35 ## he other four
36 ## The other four (incl. ASR) covary
37
38
39 ## Let's try a four factor model, merging Beauty and Hedonism
40 cfa.model2 <- 'Pragmatism =~ PQ1 + PQ2 + PQ3 + PQ4
41               Hedonism =~ HQ1 + HQ2 + HQ3 + HQ4 + Beauty
42               G =~ Goodness
43               ASR =~ indep1 + indep2 + indep3 + indep4'
44
45 cfa2 <- cfa(cfa.model2, data=HPG.fa)
46 summary(cfa2)
47 anova(cfa1, cfa2)
48 ## We should retain the four factor model
49 ## Beauty measures the same construct as Hedonism
50
51
52 cfa.model2 <- 'Pragmatism =~ PQ1 + PQ2 + PQ3 + PQ4
53               Hedonism =~ HQ1 + HQ2 + HQ3 + HQ4 + Beauty
54               G =~ Goodness
55               ASR =~ indep1 + indep2 + indep3 + indep4'
56
57
58 cfa2 <- cfa(cfa.model2, data=HPG.fa)
59 summary(cfa2)
60 fitMeasures(cfa2)
61 anova(cfa1, cfa2)
62 ## We should retain the four factor model
```

```
## Beauty measures the same construct as Hedonism

cfa.model3 <- 'Pragmatism =~ PQ1 + PQ2 + PQ3 + PQ4
               Hedonism =~ HQ1 + HQ2 + HQ3 + HQ4 + Beauty + Goodness
               ASR =~ indep1 + indep2 + indep3 + indep4'


cfa3 <- cfa(cfa.model3, data=HPG.fa)
summary(cfa3)
anova(cfa2, cfa3)



## Is Beauty and Goodness independent of Hedonism?
## This fails, most likely perhaps because the number of variables
## exceeds the number of subjects

cfa.model3 <- c(paste('PRAG =~', paste(HQ.items, collapse=' + ')),
                paste('HEDO =~', paste(HQ.items, collapse=' + ')),
                paste('BEAU =~', paste(B.items, collapse=' + ')),
                paste('GOOD =~', paste(G.items, collapse=' + ')))
cfa.model3
cfa3 <- cfa(cfa.model3, data=HPG)
summary(cfa3)
qgraph(cfa2)


## Mediator Analysis, similar to H+M(2010) ####
med.model1 <- '
# Latent Variables
PRAG =~ PQ1 + PQ2 + PQ3 + PQ4
# direct effect
PRAG ~ c*Beauty
# mediator
Goodness ~ a*Beauty
PRAG ~ b*Goodness
# indirect effect (a*b)
indirect := a*b
# direct effect
direct := c
# total effect
total := c + (a*b)'

med1 <- sem(med.model1, data=HPG.fa)
summary(med1)

## Including the intercepts
med2 <- sem(med.model1, data=HPG.fa,meanstructure=TRUE)
```

```
111 summary(med2)
112
113 #Gain model fit measures
114 fitMeasures(med2)
115 fitMeasures(med2)
116 18.147/8
```

Appendix I

Parameter estimates

In the following tables, you can find a subset of models and their parameter estimates. For each analysis, the best eight models are presented. In I.1, you can find the parameter estimates for the naive analysis, in I.2 you can find the parameter estimates for the mixed effects model analysis on the subject level and in I.3, you can find the parameter estimates for the mixed effects model analysis on the material level. As there was no theoretical reason to do so, interaction effects were not considered in the model comparisons.

TABLE I.1: Parameter estimates naive analysis

Model	Parameter estimates			
	Intercept	B	G	H
P ∼ G + H	2.61	-	0.70	-0.27
P ∼ B + G + H	2.59	-0.03	0.70	-0.24
P ∼ B + G	2.30	-0.15	0.65	-
P ∼ G	2.25	-	0.53	-
P ∼ B	3.55	0.25	-	-
P ∼ B + H	3.47	0.23	-	0.05
P ∼ H	3.38	-	-	0.29
P ∼ P	4.55	-	-	-

TABLE I.2: Parameter estimates (fixed effects) mixed-effects model (subject level

Model	Parameter estimates			
	Intercept	B	G	H
P ∼ G + H + (1 | Subj)	2.59	-	0.71	-0.28
P ∼ B + G + H + (1 | Subj)	2.57	-0.03	0.72	-0.26
P ∼ B + G + (1 | Subj)	2.26	-0.15	0.66	-
P ∼ G + (1 | Subj)	2.22	-	0.53	-
P ∼ B + (1 | Subj)	3.54	0.26	-	-
P ∼ B + H + (1 | Subj)	3.48	0.24	-	0.03
P ∼ H + (1 | Subj)	3.37	-	-	0.29
P ∼ (1 | Subj)	4.55	-	-	-

TABLE I.3: Parameter estimates (fixed effects) mixed-effects model (material level

Model	Parameter estimates			
	Intercept	B	G	H
P ∼ G + H + (1 | Website)	2.76	-	0.68	-0.30
P ∼ B + G + H + (1 | Website)	2.74	-0.04	0.70	-0.27
P ∼ B + G + (1 | Website)	2.38	0.16	0.64	-
P ∼ G + (1 | Website)	2.28	0.52	-	
P ∼ B + (1 | Website)	3.61	0.24	-	-
P ∼ B + H + (1 | Website)	3.57	0.22	-	0.03
P ∼ H + (1 | Website)	3.50	-	-	0.26
P ∼ (1 | Website)	4.55	-	-	-

Appendix J

Results PCA & EFA

J.1 Results PCA

Table J.1: Component loadings

	PC1 (H)	PC2 (P)	PC3 (I)	Communiality	Uniqueness
HQ1	0.87	0.01	0.13	0.78	0.22
HQ2	0.82	0.05	0.24	0.73	0.27
HQ3	0.88	0.13	0.11	0.80	0.20
HQ4	0.79	0.15	-0.04	0.65	0.35
PQ1	0.06	0.82	0.19	0.71	0.29
PQ2	0.22	0.86	0.13	0.80	0.20
PQ3	0.04	0.86	0.05	0.75	0.25
PQ4	0.20	0.86	0.18	0.82	0.18
Goodness	0.72	0.41	0.12	0.70	0.30
Beauty	0.88	0.07	0.09	0.78	0.22
indep1	0.14	0.11	0.83	0.72	0.28
indep2	0.21	0.10	0.89	0.84	0.16
indep3	0.05	0.07	0.90	0.81	0.19
indep4	0.04	0.22	0.61	0.43	0.57

Table J.2: SS loadings, proportion variance & cumulative variance of the components

	PC1 (H)	PC2 (P)	PC3 (I)
SS loadings	4.28	3.20	2.85
Proportion Var	0.31	0.23	0.20
Cumulative Var	0.31	0.53	0.74

J.2 Results EFA

TABLE J.3: Factor loadings

	PC1	PC2	PC3
HQ1	0.87		
HQ2	0.80		
HQ3	0.84		
HQ4	0.72		
Goodness	0.67	0.42	
Beauty	0.85		
PQ1		0.75	
PQ2		0.83	
PQ3		0.80	
PQ4		0.85	
Indep1			0.79
Indep2			0.91
Indep3			0.82
Indep4			0.45

TABLE J.4: SS loadings, proportion variance & cumulative variance of the factor loadings

	PC1 (H)	PC2 (P)	PC3 (I)
SS loadings	3.95	2.92	2.54
Proportion Var	0.28	0.21	0.18
Cumulative Var	0.28	0.49	0.67