University of Basel

Faculty of Psychology

Bachelor’s thesis presented to the Department of Psychology of the University of Basel for the degree of Bachelor of Science in Psychology

Measuring User Experience - Overview and Comparison of two Commonly Used Questionnaires

Author: Sebastian Perrig

Immatriculation number: 15-063-787

Correspondence email: s.perrig@stud.unibas.ch

Examiner: Florian Brühlmann, M.Sc.

Supervisor: Prof. Dr. Klaus Opwis

Center for Cognitive Psychology and Methodology

Submission 15.04.2018

Abstract

Since the turn of the millennium User Experience (UX) has emerged in the field of human computer interaction as an extension to the classical concept of usability. This new approach tries to grasp the user's experience as whole also considering factors beyond the usefulness of a product, such as pleasure derived from the interaction with it. Given the fairly young age of the field, scientific tools for accurate measurements are still rare and a lot of questions are brought up when it comes to the quality of the current scientific tools used in the field of UX (Bargas-Avila & Hornbæk, 2011). The goal of this Bachelor's thesis is to take a closer look at two commonly used questionnaires in the field of UX and compare them both with each other as well as with theoretical literature on scale development and questionnaire construction. In a first step, the two questionnaires VisAWI (Moshagen & Thielsch, 2010) and AttrakDiff (Hassenzahl, Burmester, & Koller, 2003) will be examined closer in terms of their development and validation process. These particular questionnaires were chosen because they both aim to measure UX in one form or another, their authors tried to construct and validate the questionnaires in a scientific way, and they are now available online for use by both researchers as well as practitioners (www.AttrakDiff.de; www.VisAWI.de). This means that even if their specific focus is different, the VisAWI's focus lies on aesthetics while the AttrakDiff focuses on the pragmatic and hedonic aspects of the user's experience, their overall goal is to make accurate scientific measurements in the field of UX. In a second step, the two questionnaires will be compared, both with each other as well as with the best practice for questionnaire and scale development and validation as presented in current scientific literature. In the end, the goal would be to have a better understanding of what it takes for a scientific instrument to be created and whether or not the two scales in question meet the criteria.

Measuring User Experience - Overview and Comparison of two Commonly Used Questionnaires

Table of Contents

Introduction

The Internet has become an everyday companion for most people. In Switzerland for example, as many as 88% of people above the age of 16 reported using the Internet at least once a week in 2016 (Bundesamt für Statistik Schweiz, 2017a). The rising popularity of the Internet can also clearly be seen. While only 51% of Europeans between the age of 15 and 74 reported using the Internet on a regular basis in 2007, the average has risen to 79% by 2016 (Bundesamt für Statistik Schweiz, 2017a). This rising popularity would therefore suggest an importance to create electronic products, such as websites and user interfaces, that are as good as they could possibly be. Users should be able to have both a satisfying as well as usable experience when interacting with electronic products. The need for precise research in what makes a product well designed is therefore given to advance our knowledge of what differentiates a good user experience from a bad one (Hassenzahl & Tractinsky, 2006). In social sciences, such as psychology, most things that are in the focus of research are often hard to grasp by simple methods of observation (DeVellis, 2017). For human computer interaction research this is no exception (Cairns, 2007). The necessity for concepts to be described by the means of theories is apparent. Since these theories are often abstract themselves, the need for precise methodology is crucial, so that the theories as well as the concepts they try to describe, can be observe in a scientific manner (DeVellis, 2017). When a new psychological field emerges, precise methodology must either be adapted from other fields of research or be created from scratch. In regards to the fairly young field of User Experience (UX), precise measurement founded in theory still seem to be scarce or knowledge about their existence seems to be lacking (Vermeeren et al., 2010). Most questionnaires used in research have not been properly validated, were self-made for the one study they were used in and are not provided to readers as part of the study even though 53% of data gathered in UX research is gathered using questionnaires (Bargas-Avila & Hornbæk, 2011). This lack of precise measurement tools should be seen as problematic when considering that a correct scientific approach seems to be one of the few reliable ways to get a grasp for latent variables that do not present themselves in directly observable behavior (DeVellis, 2017). What is it then, that makes a good research instrument precise? And are there instruments of good quality present in the field of UX? It is the goal of this Bachelor's thesis to shed some light on two currently used questionnaires from the field of UX and compare them with certain standards provided in literature when it comes to scientific scales and tests. In a first step, a common understanding of what UX is, needs to be established. After a brief explanation of UX, a few guidelines of correct questionnaire construction in accordance with best practice will be presented. Next, the two questionnaires AttrakDiff and VisAWI will be looked at, both in terms of their underlying theory as well as their scope and their process of construction and validation that has been described in scientific literature so far. Finally, a comparison of the two questionnaires with best practices in questionnaire construction should give some understanding of the quality of the scales and maybe provide some implications as to how or even if they should be used in scientific UX research.

Theoretical Background

What is UX?

As mentioned before, UX is a fairly young concept when compared with other areas of research in Psychology. Nevertheless, this does not mean that it is of less importance. With the growing amount of technology accompanying us in our everyday life, human computer interaction research, and with that, concepts such as usability and UX have become more important than ever before. Furthermore are promised to gain additional importance as technological progress continues in the way it has up until now. Certain researchers (Hassenzahl & Tractinsky, 2006; Moshagen & Thielsch, 2010) have made arguments in the past that the traditional approach in human computer interaction of only looking at usability and therefore efficiency and effectiveness does not succeed at fully capturing user demands and key factors of a satisfying interaction between a user and the product. Therefore, UX tries to go beyond just usability and aims to look at a more complete picture of a user's interaction with a product to better understand what factors are contributing to a pleasurable user experience (Hassenzahl, Diefenbach, & Göritz, 2010). Qualities that fulfill general human needs, besides direct task fulfillment, such as visual aesthetics, beauty, joy of use, stimulation, personal growth and surprise suddenly gain more importance in the overall assessment of product quality (Bargas-Avila & Hornbæk, 2011). UX therefore looks at a combination of subjective experience and feelings on the user's side in addition to the classical criteria looked at in usability (Law & van Schaik, 2010). It should therefore be of no surprise that a lot of UX research focuses on leisure products (Bargas-Avila & Hornbæk, 2011). Still, some researchers would even go as far as to say that UX should replace classic human computer interaction concepts (Hassenzahl & Tractinsky, 2006), since it fails to capture the whole picture, while others would classify UX as a subtopic of usability and therefore a key criteria for good usability to be achieved (Law & van Schaik, 2010). Summed up it can be said, that the field of UX still has a lot of questions to be answered, and precise scientific methods necessary for answering these questions still bring up many questions and issues themselves (Bargas-Avila & Hornbæk, 2011; MacDonald & Atwood, 2013; Vermeeren et al., 2010).

Best Practices of Questionnaire Construction

When it comes to the correct construction of a questionnaire, there are three main criteria that always need to be considered in wanting to guarantee highest standards of scientific research: objectivity, reliability and validity (DeVellis, 2017). All three of them, and their sub criteria, are important in order for any scientific tool to measure meaningful results in a correct manner and will therefore be summarized in the following part of this thesis in accordance with their definition in Moosbrugger and Kelava (2012) and DeVellis (2017). Since the following definitions stem from the field of test constructive theory, the word test is used. Still, they apply in the same way to our questionnaires, since they are seen as tests in this context.

Objectivity. A test is seen as objective, when its measurements of a certain trait are done without being influenced by whoever is doing the measurement or evaluating the results after measurement has been completed. In addition, clear and user-independent rules for the interpretation of the results need to be given. A differentiation is made between objectivity of application, so independence from the researcher supervising the test, and objectivity of analysis, meaning that the test results are evaluated independently from whoever is doing the evaluation. The third kind of objectivity is objectivity of interpretation, meaning that the interpretation is independent from the person looking at the results after they have been analyzed. Objectivity of application can be increased by standardizing a test and giving clear instructions on how the test should be presented, what the participant has to do, and how long it should take to complete the test. Conditions for this type of objectivity are seen as ideal, if the participant is the only source of variation during the taking of the test. It is therefore, that the usage of a computer is highly advised for increased objectivity of application. Objectivity of analysis can be achieved fairly easy by using multiple given answers to an item and not allowing open answers. Objectivity of interpretation can be increased by giving clear benchmarks and norms as a guideline for interpreting the results of a test.

Reliability. Test reliability is given if the test measures a certain trait exact and without any measurement error. It can be expressed by using the reliability coefficient, a value between 0 and 1 that should not be under .70 for any test to have good reliability. In theory, all measurements of variance are made up of the true variance, represented by the reliability, and the measurement error, referred to as unreliability. There are different measurements for showing reliability, one of the most used being internal consistency represented by Cronbach's alpha, a reliability coefficient representing the correlation between all the different items of a test with each other, first presented in Cronbach (1951). High correlation of the different test items with each other are reflected in a high coefficient value and therefore speak for high internal consistency and an overall more reliable test. Important for this process to work is that all items included in the calculation of Cronbach's alpha measure the same trait. What has to be kept in mind though when using Cronbach's alpha is its dependency on the amount of items. Cortina (1993) was able to demonstrate, that when increasing the items of a fictional test, coefficient values for Cronbach's alpha reached values above .70 after 14 or more items were included in the test. This was the case even though the correlation between the items was only at .30. With higher correlations between the items, the coefficient value increased even further. Cronbach's alpha can therefore be seen as a value for measuring internal consistency, but not directly for one dimensionality of the data measured by the test according to Cortina (1993).

Validity. A test is of high validity, if it measures the trait it aims to measure and not something else (Moosbrugger & Kelava, 2012). While reliability and objectivity are important because they lay the foundation for high validity by creating measurement accuracy, it is high validity itself that makes the results of a test useful. High validity is key in being able to transfer test results from a test situation to behavior outside of the artificial environment. Differentiations are made between several kinds of validity. Content validity refers to the extend to which the test captures the trait it aims to measure in a representative way. This criterion is not evaluated through direct measurement but rather through assessment, mostly done by experts. Face validity is given, if validity is present in the eyes of a layperson looking at the test. Construct validity is given, when a measured construct relates to other, scientifically measured constructs, in a way that would be expected from the theory behind the test. Exploratory factor analysis can be used to gain measurements for construct validity. This is done by assessing the structure of the test items and calculating both the convergent as well as divergent validities. Convergent validity represents to what extent the test results correlate with other tests measuring for same or very similar traits. The higher the correlations, the better the convergent validity. Divergent validity on the other hand measures to what extent the test correlates with other tests capturing traits different from the trait measured by the test that's being assessed. Here, low correlations are desirable. A combination of both convergent and divergent validity give important information about the overall validity of the test. Criterion-related validity asks, whether or not a transfer can be made from a test subject's score to a certain criterion observed. This is usually measured by looking at a real criterion familiar with the trait measured by the test. The higher the correlation between test score and criterion measured, the better the criterion-related validity.

Factor Analysis and Factor Rotation. One key procedure in the development and validation of scientific scales and tests is factor analysis, which is explained in DeVellis (2017) as follows: Its importance stems mainly from the concept, that a lot of times, more than one latent variable lies at the source of variation shown by the items of a scale. What can therefore be done, through the process of factor analysis, is finding the amount of factors underlying the variation that is able to explain it the best. Once factors are identified through analysis, looking at the content of the items making up the factors found can often give insight into the underlying latent variable represented by the factor. In addition, factor analysis can also be key to eliminating items that do not serve any purpose in measuring the latent variable by removing those items without any connection to one of the factors extracted in the analysis process. Factor analysis itself though does not provide anything more than the appropriate number of factors to extract. Another important step that has to be done to further understand the nature of a data frame is factor rotation (DeVellis, 2017). This refers to the practice of transforming data so it is represented to us in a way that is more understandable. The items forming the data are not changed in this process and neither are the relationships between them. What makes this process useful though is the finding of item clusters through a change of perspective, allowing us to identify common factors representing latent variables in a set of items that can then be key to further interpretation of the data as a whole. Still, the process of factor analysis brings risks of error with it. Many decisions have to be made when going through the process of conducting a factor analysis, such as the choice of factor rotation method or the cutoff point for factor loadings, and these decisions can cause errors when not going in the right direction, which is something that seems to happen rather often in human computer interaction research (Howard, 2016). It therefore should be clearly noted, that factor analysis is not without flaws and should be used in a precise and careful manner.

Goal of this Thesis

Now that certain guidelines on what makes a good test have has been established, the next part of this thesis will focus on the two questionnaires, the AttrakDiff and the VisAWI. The goal of this process will be to have a clear understanding of the construction and validation processes for the two scales so that in the end, the following questions can hopefully be answered: To what extend have the two questionnaires AttrakDiff and VisAWI been constructed and validated in accordance with best theoretical practice? What are the implications of this development process for their practical usage in the field of UX research?

About the AttrakDiff – Explanation and Development

The AttrakDiff is a semantic differential build up of 28 items rating a product on both hedonic as well as pragmatic qualities from the perspective of the user (Hassenzahl et al., 2003). The basis for the importance of pragmatic as well as hedonic factors in user experience stems from Hassenzahl's theory brought up and later on expanded in his papers such as Diefenbach, Kolb, and Hassenzahl (2014). As of writing this thesis, the AttrakDiff has around 672 citations [1].

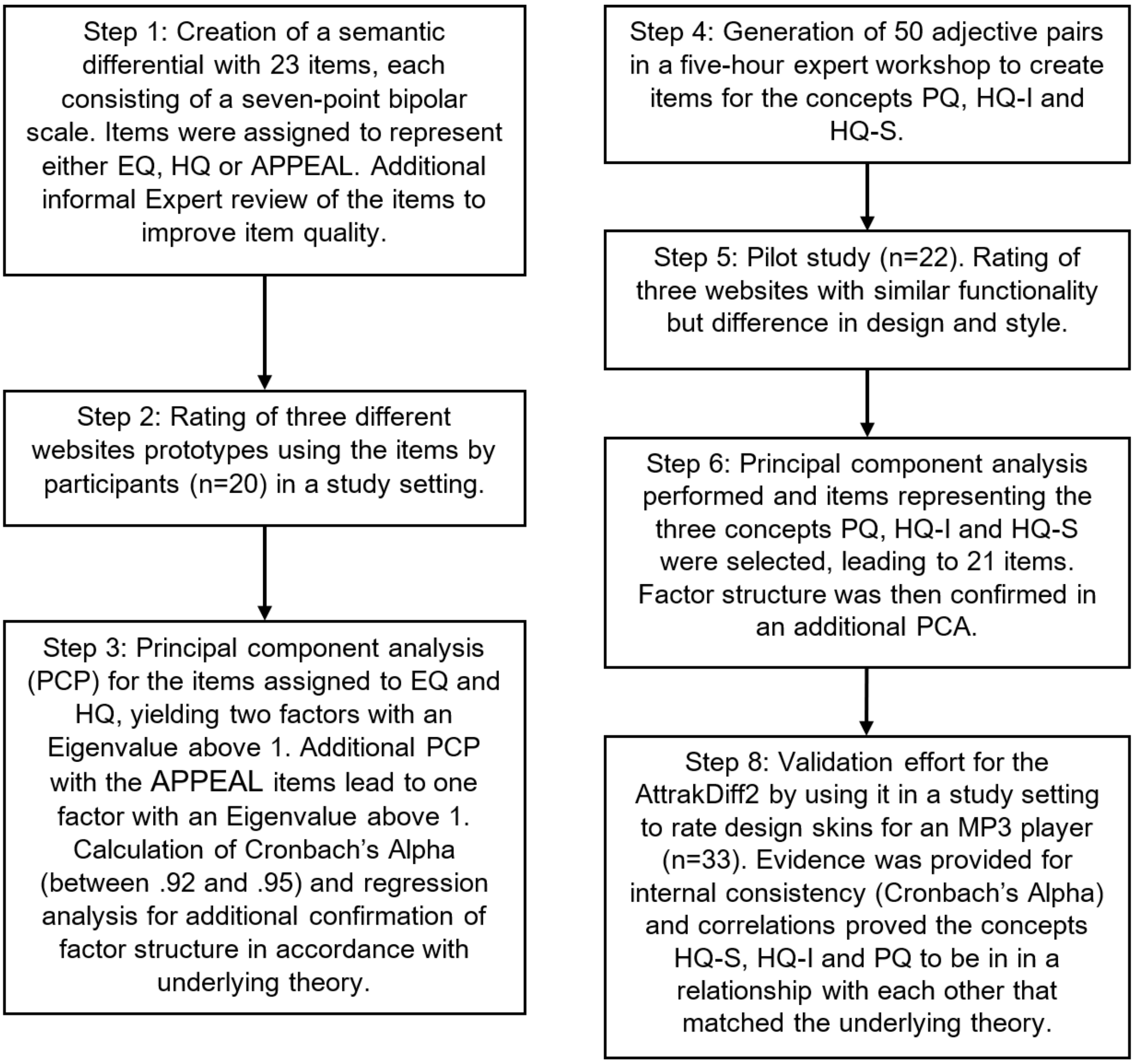
Underlying Theory for the AttrakDiff

At the basis of the AttrakDiff questionnaire lies a theory explained as follows in Hassenzahl, Platz, Burmester, and Lehner (2000). In their paper, Hassenzahl and his colleagues aimed to create a better understand of factors that contribute to an enjoyable user experience when interacting with a product. They mention two main qualities that determine a user's feelings towards a piece of software, hedonic and ergonomic qualities, and argue that these two qualities need to be considered in order for the limited concept of usability to be expanded. Ergonomic quality (EQ) takes up the aspects usually associated with classical usability, namely things such as efficiency, effectiveness and other task related aspects relevant for user interaction. Hedonic

[1] 11.04.2018 on Google Scholar for the original AttrakDiff2 paper (Hassenzahl et al., 2003)

quality (HQ) on the other hand has no direct connection with the execution of a task by the user but rather has importance in itself. Aspects of HQ such as beauty, originality and innovativeness influence the overall perception and enjoyment a user has when interacting with a product. Moreover, Hassenzahl and his colleagues argue that it is not the objective level of EQ and HQ for a product that matters but more the subjective qualities perceived by the user. They therefore aimed to create a way to measure both EQ as well as HQ from the subjective viewpoint of the user. A few years after the first paper about the AttrakDiff, a second version of the questionnaire was created and in that process, the theory was expanded further. In Hassenzahl et al. (2003) a further distinction for hedonic quality was added. In addition to the already mentioned ergonomic quality and the overall attractiveness of a product (ATT), the new version of the AttrakDiff was created to make a further distinction between two sub-concepts of HQ. One of the newly added concepts was user stimulation (HQ-S). This describes the amount of personal development, such as acquiring new knowledge and abilities, that a user hopes to gain from use of a product. The second newly added concept was user identity (HQ-I), with which the phenomena tries to be captured, that users also try to express part of their desired identity by choosing one product over another. In addition, the term ergonomic quality (EQ) used in Hassenzahl et al. (2000) was replaced by the term pragmatic quality (PQ). Even though further expansions of the theory were made in later years, they shall not be further discussed at this point since the focus of this paper should lie on the construction and validation of the original questionnaires AttrakDiff and AttrakDiff2. This is mainly due to the fact, that the AttrakDiff2 is still used in research in the version described in Hassenzahl et al. (2003).

Construction and Validation of the AttrakDiff



Step 1: Creation of a semantic differential with 23 items, each consisting of a seven-point bipolar scale. Items were assigned to represent either EQ, HQ or APPEAL. Additional informal Expert review of the items to improve item quality.

Step 2: Rating of three different websites prototypes using the items by participants (n=20) in a study setting.

Step 3: Principal component analysis (PCP) for the items assigned to EQ and HQ, yielding two factors with an Eigenvalue above 1. Additional PCP with the APPEAL items lead to one factor with an Eigenvalue above 1. Calculation of Cronbach's Alpha (between .92 and .95) and regression analysis for additional confirmation of factor structure in accordance with underlying theory.

Step 4: Generation of 50 adjective pairs in a five-hour expert workshop to create items for the concepts PQ, HQ-I and HQ-S.

Step 5: Pilot study (n=22). Rating of three websites with similar functionality but difference in design and style.

Step 6: Principal component analysis performed and items representing the three concepts PQ, HQ-I and HQ-S were selected, leading to 21 items. Factor structure was then confirmed in an additional PCA.

Step 8: Validation effort for the AttrakDiff2 by using it in a study setting to rate design skins for an MP3 player (n=33). Evidence was provided for internal consistency (Cronbach's Alpha) and correlations proved the concepts HQ-S, HQ-I and PQ to be in in a relationship with each other that matched the underlying theory.

Figure 1. Flowchart summarizing the process of construction and validation for the AttrakDiff (steps 1 to 3) and the AttrakDiff2 (steps 4 to 8) as reported in Hassenzahl et al. (2000) and Hassenzahl et al. (2003).

First construction of the AttrakDiff was done as part of Hassenzahl et al. (2000) and first efforts to validate the questionnaire were also undertaken as part of that paper. In Hassenzahl et al. (2003) the more fleshed out AttrakDiff2 was created and validated at the same time. Since 2003, the AttrakDiff2 is accessible as an online tool for both researchers as well as practitioners, with the goal of not only making the AttrakDiff2 attractive for use but also to gather more data on how it is performing in different usage contexts (Hassenzahl, Koller, & Burmester, 2008). The process of validation and construction the AttrakDiff as described in these two papers (Hassenzahl et al., 2003, 2000) will now be summarized and is also visualized in Figure 1.

Construction of the AttrakDiff1. For the original creation of the AttrakDiff in Hassenzahl et al. (2000), the theory mentioned before was taken as a basis. The researchers then aimed to find out, whether or not hedonic and ergonomic quality could be measured as concepts independently from each other and to what extend the two factors would contribute to the overall perception of attractiveness of a website and therefore its appeal (APPEAL). In a first step, a semantic differential with 23 items was created, each item being made up of a seven-point scale with two bipolar anchors. Details about the creation process of these items were not further described. These 23 items were selected to either represent HQ, EQ or APPEAL. In addition to the selection process of the items, an informal expert review of the items was conducted to further improve item quality. The specifics of this review were not described in the paper. The items where then used in a first study were 20 participants had to rate website prototypes with them. After results had been gathered, a principal component analysis (PCA) looking at the items for HQ and EQ was performed leading to the extraction of two factors with Eigenvalues above 1, together explaining 68% of variance. These factors were assumed to represent EQ and HQ. A PCA for the APPEAL items only identified one factor with an Eigenvalue above 1. Cronbach's alpha values for the three scales, both measured before as well as after interaction with the software, were between .92 and .95. Two regression analysis revealed no interaction between HQ and EQ but an equal contribution of the two factors to the value of APPEAL. Overall, the authors saw their results as supportive of their theory of hedonic and pragmatic quality as independent factors both influencing the overall appeal of software.

Construction of the AttrakDiff2. A few years later, in Hassenzahl et al. (2003), a new version of the AttrakDiff, the AttrakDiff2 was presented with the goal of being able to grasp hedonic as well as pragmatic qualities even better than its predecessor. In addition, the aforementioned expansion of the underlying theory, now also considering stimulation (HQ-S) and identity (HQ-I) was to be implemented into the AttrakDiff2. The original terminology of ergonomic quality (EQ) was replaced by pragmatic quality (PQ) and instead of overall appeal (APPEAL), attractiveness (ATT) was used as a new term. In a first construction step, a five hour expert workshop was conducted to gather possible items consisting of adjective pairs. Ratings for these items were then performed to evaluate their suitability for the new version of the AttrakDiff. This lead to 50 items, in addition to the seven original APPEAL items from the AttrakDiff1, to form the ground work for further empirical construction. Next, a pilot study with 22 participants recruited through newspaper adds was conducted. Three websites were chosen, all with similar functionality but great difference in design and style of interaction. The new version of the AttrakDiff was used by the participants to rate the three websites. After rating data from the participants had been gathered, principal component analysis (PCA) and varimax rotation for the 50 items from the expert workshop was conducted with the goal of precisely extracting the three factors HQ-S, HQ-I and PQ. Items were chosen based on their loadings so that each factor would have at least six items. Through this process, 21 items were identified. PCA was then performed again, this time not with regard to their affiliation with one of the factors HQ-S, HQ-I or PQ, but with the criterion of an Eigenwert above 1 for extracted factors. Three factors were found through this process that overall could explain 72% of variance. HQ-S was able to explain 29% of variance with item loadings between .758 and .900. HQ-I could explain 23% of variance with loadings of .684 to .831. PQ explained 20% with loadings from .642 to .685. The expected structure from the theoretical background was in accordance with the structure shown by the data. To further validate the questionnaire, a second study was made, this time with 33 psychology students. Participants had to use a software for playing MP3 files and had to rate different versions of the software using the AttrakDiff2. Only the design of the software was changed using skins while the functionality was kept the same. Internal consistency for the different subscales was calculated using Cronbach's alpha, delivering good to very good results (HQ-S: .76-.90; HQ-I: .73-.83; PQ: .83-.85). Low correlations between HQ and PQ (.18 for PQ and HQ-S; -.13 for PQ and HQ-I) again showed the two factors to be independent from each other. HQ-S and HQ-I correlated rather high with a value of .55 as would be expected from the theory. Due to a lack of a priori Hypothesis, statements about construct validity could not be made. Overall, the authors deem the quality of the AttrakDiff2 to be acceptable and further validation not necessary. In their eyes, future focus should lie on research regarding the base theory. The AttrakDiff2 itself should be used as a research tool and not as a fully developed method of evaluation. Whether or not this truly is the case shall be discussed later.

Using the AttrakDiff - Instructions and Examples

Since 2003 the AttrakDiff2 is not only available in its paper form but also digitally over the website www.AttrakDiff.de (Hassenzahl et al., 2008). During the first four years of its online availability, around 2300 evaluations using the online AttrakDiff2 had been made as part of 302 projects. Most of theses projects were praxis orientated and aimed to evaluate interactive products. The website offers the possibility of one time evaluations (80% of projects), before-after comparisons (5%) and comparisons of two products (15%). Projects can contain a maximum of 20 participants and can be active for the duration of up to three months.

Examples of use in Research. A few examples of the AttrakDiff's use in research will be shortly explained at this point to get a better grasp of the AttrakDiff's potential. In order for users to become more motivated in the use of renewable energy sources in modern electricity grids, a game-like system was developed to create long term commitment (Gnauk, Dannecker, & Hahmann, 2012). To test that system's appeal, both the System Usability Scale (Brooke, 1996) for measuring usability and the AttrakDiff2 was used. Results produced by the two questionnaires were comparable in nature and in accordance with feedback given by the users. Another example from peer reviewed literature was a study (Gebhardt, Pick, Oster, Hentschel, & Kuhlen, 2014), where the AttrakDiff2 was used alongside the System Usability Scale (Brooke, 1996) and the User Experience Questionnaire (Laugwitz, Held, & Schrepp, 2008) to get measurements of the user's experience with two different menu systems for controlling a virtual reality software. One of the two menus used in the study was a newly created mobile phone menu, while the other one was an already established extended pie menu. All three scales came to similar results in regards to UX measurements. In Grubert, Pahud, Grasset, Schmalstieg, and Seichter (2015) the AttrakDiff2 was used for measuring pragmatic as well as hedonic qualities of a new interface in comparison with an already established one on a small handheld device. Notably, the pragmatic quality items of the AttrakDiff2 were used alongside the NASA-TLX (Hart & Staveland, 1988) and both scales could deliver comparable, significant results. In Thanh Vi, Hornbæk, and Subramanian (2017), participants inside an fMRI scanner were shown websites with different usability and aesthetics ratings to identify brain areas related to these website design aspects. Items from the AttrakDiff2 were used for the stimuli selection of those websites, that were to be presented to the fMRI-participants.

About the VisAWI – Explanation and Development

In this section, the VisAWI (Moshagen & Thielsch, 2010) will be looked at. The VisAWI works with 18 Likert scale items to get a rating for the aesthetics of a website (Thielsch & Moshagen, 2011). With only 297 citations[2] the VisAWI is less popular than the AttrakDiff, but the setting in which it was developed as well as the way it is distributed is similar. Both were developed in a German setting and are now distributed online over their own respective Websites in corporation with the User Interface Design GmbH (www.AttrakDiff.de; www.VisAWI.de). They also both try to measure important aspects of UX, which is a reason why they can be seen as comparable and are available for use in combination over their websites. Also, the fact that the AttrakDiff is about seven years older than the VisAWI might play a role for the difference in citations. In addition, aesthetics has been shown to be strongly correlated with perceived usability (Tractinsky, Katz, & Ikar, 2000). First impressions of a website also strongly rely on judgments of aesthetics, which are made as quick as 50 milliseconds (Lindgaard, Fernandes, Dudek, & Brown, 2006) and remain highly stable over time (Tractinsky, Cokhavi, Kirschenbaum, & Sharfi, 2006). This therefore demonstrates the

[2] 04.04.2018 on Google Scholar for the original VisAWI paper (Moshagen & Thielsch, 2010)

importance of a good measurement tool for aesthetics inside the field of UX, making the VisAWI interesting for this thesis as well as for UX research as a whole.

Underlying Theory for the VisAWI

The importance of aesthetics for UX has been argued about in Moshagen and Thielsch (2010) and lays the theoretical groundwork for the VisAWI. In their paper, the authors argue that visual aesthetics are a key influencer to concepts such as perceived usability, user satisfaction and pleasurable usage experience. Because of this, they wanted to have a clear definition of what visual aesthetics are and after that, create an adequate tool for scientific measurement. To understand aesthetics, Moshagen and Thielsch (2010) looked at common definitions of beauty and settled on an interactionist perspective as suggested in philosophy by George Santayana (1955). This viewpoint sees beauty as a result of an interaction between the perceiver's characteristics and the perceived object's properties. Beauty therefore would evoke pleasurable feelings in its perceiver and would in its nature be objective in that these feelings are connected to the perceived object itself. Reasoning does not play a role at this moment of perception. Finally, it can be seen as intrinsic, since it does not serve any direct expected utility. Because this interactionist viewpoint of beauty resembles the viewpoint of aesthetics in scientific research (Reber, Schwarz, & Winkielman, 2004; Solso, 2003), the two were seen as equal for the purpose of the paper. Furthermore, the authors made the argument, based on aesthetics literature, that aesthetic response is highly dependent on how fluently a perceiver is able to process an object. In the process of creating the VisAWI, four aspects of visual aesthetics relevant for a website were identified. These four also reflected the principles of aesthetics brought up in the definitions chosen by the authors. Simplicity is seen as all those aspects that can make perception easier for the user (e.g. clarity and orderliness). Diversity is important for website aesthetics, because it can give the product a feeling of visual richness (e.g. dynamics, variety and creativity). The use of individual colors as well as their composition make up the third aspect, described as colorfulness. The last important aspect, craftsmanship, summarizes to what extend a product was created in a skillful way with careful design and in accordance with modern standards.

Construction and Validation of the VisAWI

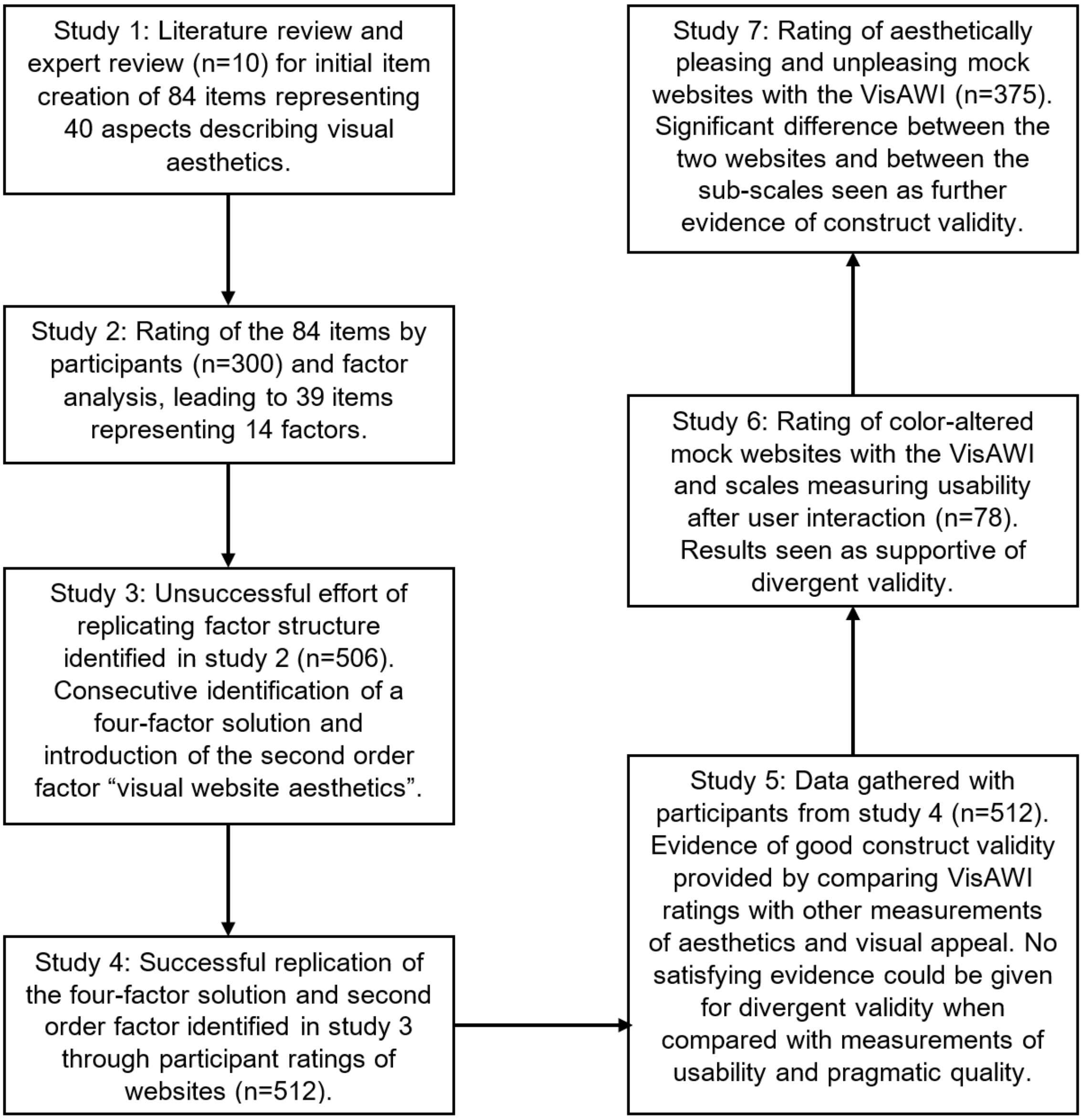

Study 1: Literature review and expert review (n=10) for initial item creation of 84 items representing 40 aspects describing visual aesthetics.
Study 2: Rating of the 84 items by participants (n=300) and factor analysis, leading to 39 items representing 14 factors.
Study 3: Unsuccessful effort of replicating factor structure identified in study 2 (n=506). Consecutive identification of a four-factor solution and introduction of the second order factor “visual website aesthetics”.
Study 4: Successful replication of the four-factor solution and second order factor identified in study 3 through participant ratings of websites (n=512).
Study 5: Data gathered with participants from study 4 (n=512). Evidence of good construct validity provided by comparing VisAWI ratings with other measurements of aesthetics and visual appeal. No satisfying evidence could be given for divergent validity when compared with measurements of usability and pragmatic quality.
Study 6: Rating of color-altered mock websites with the VisAWI and scales measuring usability after user interaction (n=78). Results seen as supportive of divergent validity.
Study 7: Rating of aesthetically pleasing and unpleasing mock websites with the VisAWI (n=375). Significant difference between the two websites and between the sub-scales seen as further evidence of construct validity.

Figure 2. Flowchart summarizing the process of construction (studies 1 to 4) and validation (studies 5 to 7) for the VisAWI as reported in Moshagen and Thielsch (2010).

In Moshagen and Thielsch (2010), a lot of focus is set on the importance of validating a newly developed scale. It therefore should be expected of the paper to present a clear process of how the scale was constructed and validated. It is therefore no surprise, that the VisAWI was developed in a process of seven studies, four for construction and three for validity. These studies are described in the following section and are also summarized in Figure 2.

Study 1. In the first study, aspects describing visual aesthetics were identified by reviewing empirical and theoretical literature sources. These aspects were then rated by 256 web users in accordance with their importance for website aesthetics, also giving them the possibility of suggesting additional aspects if they thought any were lacking. After that, 10 experts undertook a think aloud method of website evaluation to identify aspects used by them. These aspects were then compared with the original aspects from the literature review and web user rating, as well as with the criterion of only rating a single website and not a cluster of several. This process left the researchers with a final total of 40 aspects. For the use of item generation, theses aspects were then sorted into 12 broad domains. Items were then created in accordance with the findings from the previously described interview, creating more items for those aspects deemed of higher importance by the web users interviewed. After eliminating items of low quality, decided upon by ratings from four experts and three laypeople, a total of 84 items were created, with at least five per domain, half of them reverse coded.

Study 2. In the second study made as part of the construction process, 300 participants had to use the 84 items to rate websites in a language they could not understand, so the focus would be on the design itself and not the content of the websites. The subjects also had the possibility to flag items they thought to be of questionable quality (e.g. ambiguous). In a next step, exploratory factor analysis was used, leading to 14 factors with an Eigenvalue greater than 1. To avoid the risk of too many factors, an additional criterion was applied to create clearer factors. Through parallel analysis of 1000 randomly generated data sets, a solution of six factors was deemed fitting with 53% of variance explained. Inter-correlations for the factors ranged from .14 to .46. After keeping all items that showed primary factor loadings above .4 and no significant secondary loadings, a final set of 39 items was kept for the next study.

Study 3. The third study aimed to replicate the factorial structure found beforehand and to do further refinement as well as shorten the scales by means of confirmatory factorial analysis (CFA). For this, the same methods as in study 2 were used, this time on a sample of 506 volunteers recruited over the Internet. This time though, the new 39 item-version of the questionnaire was used and the option of flagging items was not implemented anymore. Because the results of the CFA were only acceptable and therefore non-satisfactory, a four factor solution was considered. The new structure was able to explain 67% of variance leaving us with the four factors explained in the theory about the VisAWI beforehand (simplicity, diversity, colorfulness and craftsmanship). This was also achieved by combining some of the factors from the previous six factor version that showed high correlations with each other suggesting overlapping (r=.97 for two of the factors and r=.88 for two others). Because of high correlations between the four remaining factors (.60<r<.74) a second order factor "visual website aesthetics" was introduced. After dropping some items because of high secondary loadings and a few more due to low primary loading to improve model fit, 18 items remained. Confirmatory factor analysis showed this new version to have primary loadings above .60 without substantial secondary loadings, therefore having satisfyingly better results than the previous six factor version.

Study 4. In the fourth and final study of the construction process, an identical procedure as in the previous studies was used. This time though, the 512 volunteer participants had to rate a randomly drawn website out of a pool of 42 popular institutional and corporate German websites. This meant that the website choice this time was more natural due to participants being able to understand the website language. CFA showed that primary loading magnitudes ranged from .63 to .89, without any secondary loadings of substance. In addition, the four facets seem to equally represent the general factor "visual websites aesthetics". Cronbach's alpha Values for the four scales ranged from .85 to .89 and for the total score a value of .94 was calculated.

Study 5. The fifth study aimed to calculate the construct validity for the now constructed VisAWI. For this, the same sample as in study 4 evaluated the same set of websites. In addition to the VisAWI, several other questionnaires commonly used in UX and neighboring fields were used. For convergent validity, measurements from a scale assessing perceived visual aesthetics in a website context were collected (Lavie & Tractinsky, 2004). In addition, the scales measuring pragmatic quality and overall attractiveness from the AttrakDiff (Hassenzahl et al., 2003) were used. For perceived website usability, a scale by Flavián, Guinalíu, and Gurrea (2006) was used and the overall quality of content was measured by items from Thielsch (2008). The order in which the questionnaires, as well as their items were presented, was randomized. To demonstrate convergent validity, the VisAWI's values were compared with other instruments assessing aesthetics and visual appeal. Correlations between the values were high, suggesting good convergent validity. For the demonstration of divergent validity, tools for the measurement of usability and pragmatic quality were correlated with the results from the VisAWI. This lead to results that could not be seen as in support of divergent validity. A highly significant MANOVA showed that the VisAWI made significant distinctions between different websites of unequal levels of aesthetics. This was seen as prove for the discriminative validity of the VisAWI. Furthermore, ANOVAs demonstrated significant differences among the sub-scales in regard to different websites and the effect size between the lowest and the highest graded website showed a very large effect (d=2.0). To demonstrate concurrent validity, correlations between the VisAWI scores and the participant's intent to revisit the websites were done, showing positive results.

Study 6. In the sixth study, the doubts about divergent validity raised in the previous study were addressed again. The websites rated this time by the 78 volunteer participants were one of two mock websites that were only altered in terms of colorfulness to either be aesthetically pleasing or distorted. Since the authors figured that perceived usability and website aesthetics might have had such a high correlation in the previous study because participants did not interact with the website but only looked at it, task were given to the participants to complete before rating the website in terms of both usability as well as with the VisAWI. Correlations between perceived usability and the VisAWI scores this time were much weaker and only significant for the sub scales simplicity and craftsmanship as well as for the overall general factor.

Therefore, further evidence for the divergent validity of the VisAWI was provided in the author's opinion. Again using a MANOVA, significant differences could be shown between the two conditions and ANOVAS showed that the difference was only present on the colorfulness sub-scale, providing additional evidence for the VisAWI's construct validity.

Study 7. To further demonstrate construct validity, one last study was done. Aesthetically pleasing and aesthetically distorted versions of a mock website with an otherwise unaltered structure were created for comparison using the final version of the VisAWI. A total of 375 participants were assigned to one of the two website conditions. A MANOVA was able to show significant differences between the two conditions and several ANOVAs demonstrated significant differences in the sub-scales, offering further support for the construct validity of the VisAWI.

Additional Remarks. Overall, the authors saw the VisAWI as a good tool to make subjective measurements of website aesthetics from the perspective of the user. Three years after the creation of the original VisAWI, a shorter version consisting of only four items taken out of the original questionnaire was created (Moshagen & Thielsch, 2013). The focus of this paper will though remain on the original VisAWI, since the VisAWI-S is not an update, but rather a different version of the VisAWI providing comparable results for use in areas where the full VisAWI would not be practical time-wise (see appendix B for VisAWI and VisAWI-S items).

Using the VisAWI - Instructions and Examples

Just like the AttrakDiff, the VisAWI can also be used over its own respective website, www.VisAWI.de. The principle for use is the same as with the AttrakDiff and both can even be used in combination with each other. For a pen and paper use of the questionnaire, a clear manual with instructions is given (Thielsch & Moshagen, 2014). The Manual gives a short summary of the theoretical groundworks for the VisAWI, summarizes the construction and validation process and delivers the user with bench mark and norm tables for better interpretation of the results. If needed, the four subscales of the VisAWI can also be used on their own (Thielsch & Moshagen, 2011).

Examples of use in Research. One example where the VisAWI was used in a research setting can be seen in a study trying to examine the relationship of usability and aesthetics for websites, where users do not already have a clear model of what the website should be like (Stojmenovic, Pilgrim, & Lindgaard, 2014). For this purpose, city council websites were first marked as fitting the criteria and then graded using the VisAWI-S and the System Usability Scale (Brooke, 1996), showing that the VisAWI-S can be used in different contexts. In addition, the results showed significant correlations between the usability measured with the System Usability Scale, and the visual appeal measured with the VisAWI-S. Since the VisAWI-S is made up of items from the VisAWI (see appendix B) and is intended to present similar results to the full VisAWI in settings with less time (Moshagen & Thielsch, 2013), results achieved with the VisAWI-S should tell us at least partially about the way the VisAWI would perform in such a situation. In Linares-Vásquez et al. (2015), the items measuring colorfulness were taken out from the VisAWI and were presented as four-point Likert scales to rate different graphical user interfaces. The interfaces were adjusted in their color composition to cause lower energy consumptions and the attractiveness of the low-energy interfaces was indicated through the user's feedback using the VisAWI items. Another example of usage for the VisAWI was a study that tried to create an integrated measurement scale for assessment of smartphone user interfaces in terms of their simplicity (Choi & Lee, 2012). Three items were drawn from the VisAWI to asses aesthetic simplicity and were rephrased to fit mobile phones before being integrated into the newly created scale. One last study used the VisAWI alongside the System Usability Scale (Brooke, 1996) to get a better understanding of the relationship between usability and aesthetics in a website context (Stojmenovic et al., 2014). The results suggested positive correlations between the two concepts both pre- and post-use.

Discussion

Comparison of AttrakDiff & VisAWI with Best Practices of Construction

Now that both the AttrakDiff as well as the VisAWI have been fully looked at in terms of their construction and validation processes, as well as in a few examples of use, they shall be looked at closer in regards to the standards of scientific questionnaire construction described beforehand.

Objectivity. First, the question of objectivity needs to be considered. Since both tests are standardized in their questions and do not offer the possibility of open answers, the criterion of objectivity of application seems to be given at least to a certain extend. A strong point for both questionnaires is the use of a website, not only leading to easy distribution of the questionnaire but also increasing objectivity of application, analysis and interpretation through automating the whole process. Whether or not other possible sources of error might come up when performing studies online would although be a whole other field of research to get into. Since both questionnaires are available over the User Interface GmbH, a usage of the VisAWI website will also recommend using the AttrakDiff to the user and vice versa, therefore making the two interlinked to a certain extend. Still, objectivity might be hard to achieve when all that is present for the paper version of the AttrakDiff seem to be the 23 items without any manual or additional information on how to use the AttrakDiff. This might also be one reason why the AttrakDiff was recommended to not be seen as a full evaluation method by its own authors, bringing up the question of whether or not it should even be used as a questionnaire in the way it seems to be used in the research examples provided beforehand. In comparison, the VisAWI has a full manual, mainly focused on the explanation of both the underlying theory as well as a quick summary of the construction process (Thielsch & Moshagen, 2014). In addition, the manual provides clear instructions on how to use the VisAWI paper version, alongside with instructions on how to evaluate the results. For that purpose, benchmark tables that were done as part of later papers following the original construction of the VisAWI are provided. This would therefore suggest that objectivity of interpretation might be higher for the

VisAWI, at least when compared with the AttrakDiff's paper version. This would have to be confirmed though, for example by using the two questionnaires alongside each other in a scientific study.

Reliability. In terms of reliability, judgments can mainly be made by looking at the reliability coefficients provided at several points in the papers on both the AttrakDiff and the VisAWI. As a reminder, high internal consistency and therefore reliability is presented when the Cronbach's alpha coefficient reaches values of .70 and higher. For the AttrakDiff2, coefficient values range from .73 to .90 therefore demonstrating that internal consistency seems to be given. For the VisAWI, Cronbach's alpha values were calculated for both the sub scales as well as for the overall score. The sub scale coefficients ranged from .85 to .89 while the total score demonstrated a value of .94. Therefore internal consistency for the VisAWI can also be seen as present. Still, other methods of measuring reliability, such as split-half or retest methods (Moosbrugger & Kelava, 2012), would be useful in addition to Cronbach's alpha, to get a more consolidated grasp of the two questionnaire's reliability.

Validity. Next, the validity of the two questionnaires should be considered. As established beforehand, content validity can mainly be improved by getting expert opinions on a test and its items. For the first AttrakDiff, this was arguably done in form of the original informal expert review. When construction of the AttrakDiff2 was done, a more extensive method was chosen by having a five hour expert workshop. Therefore, expert opinions were used for the initial choosing of the items making up the AttrakDiff2, most likely leading to content validity. For the VisAWI, experts were included in the construction process on several occasions. Experts were used for the original gathering of the test items and also later for judging item quality. This would lead us to expect content validity to also be present for the VisAWI. Face validity was only looked at when constructing the VisAWI and not in the process of constructing the AttrakDiff. What has to be said though at this point, is that face validity is not of very high importance for a questionnaire's quality (DeVellis, 2017). Construct validity on the other hand should be taken into detailed consideration. Fortunately, exploratory factor analysis was done in the process of constructing the VisAWI. Through this process, high convergent validity was proven to be present for this questionnaire. Evidence for divergent validity was also provided as part of the construction process, although weaker than the evidence provided for convergent validity. This might also point out to the viewpoint of some researchers, that aesthetics influence other parts of usability to an extend, and therefore measurements of divergent validity might be influenced by that relationship. Besides the evidence provided in Moshagen and Thielsch (2010) for construct validity, examples of usage for the VisAWI only provide us partially with evidence of convergent or divergent validity. Most of the times, when used alongside other scales, only certain items of the VisAWI or the VisAWI-S would be used, but rarely the full VisAWI. While this might tell us something about practical implications of use for the VisAWI, mainly that it might be considered too long by researchers, it does not help us much in terms of validity assessment. Still, the original paper of the VisAWI already provided clear evidence of the VisAWI's construct validity. For the first version of the AttrakDiff, principal component analysis (PCA) was conducted for the two factors EQ and HQ as well as for APPEAL, but no exploratory factor analysis was done. The question of construct validity remained unclear for this version of the AttrakDiff. The AttrakDiff2 was also only subject to PCA but not to exploratory factor analysis, making a judgment of construct validity on this basis impossible. Even the authors themselves had to admit that: "Since no independent expert review or any other type of evaluation had been undertaken prior, results could strictly speaking not be interpreted as evidence of construct validity" (Hassenzahl et al., 2003, p.8, translated from German). Evidence for convergent validity was presented to a certain extent, when considering the previously presented examples from scientific research, where the AttrakDiff was used. Whenever the AttrakDiff was used alongside other questionnaires, measuring similar traits, results were comparable and no problems were reported by the authors in using the AttrakDiff. In order for a clearer understanding of convergent validity, as well as divergent validity, exploratory factor analysis would be desirable. In order for us to judge criterion validity, future research would probably have to be done to see whether or not websites with high levels in the questionnaires are also seen as more positive outside of the experimental setting, for example by having a good personal image or being successful at fulfilling their purpose.

Additional Remarks. In regards to the underlying theories at the basis of the two questionnaires, it has to be noted that the VisAWI focuses on theories already established by several works of scientific literature before it such as Reber et al. (2004) and Solso (2003). The AttrakDiff on the other hand relies on a theory created by the authors themselves. It also seems that the main goal of the authors of the VisAWI was to provide a "precise operational definition and to develop a new measure of perceived visual aesthetics of websites" (Moshagen & Thielsch, 2010, p.1). The authors of the AttrakDiff on the other hand clearly said that while they do hope that the AttrakDiff will be further used both in research as well as by practitioners, focus should not be laid on "further validation of the questionnaire but on better understanding the underlying model" (Hassenzahl et al., 2003, p.8, translated from German). Another important point to consider are the samples used for the scientific construction and validation of the two questionnaires. On its own, the first validation of the AttrakDiff (Hassenzahl et al., 2000) would most likely not have been enough as a foundation for a reliable research tool. The fairly small sample size of only 20 participants, all being workers in the same company, would not satisfy demands for transferability to a general population. For the creation of the AttrakDiff2 (Hassenzahl et al., 2003), the participant choice might also be criticized for both studies done. While the age range (from 23 to 59) might be representative for a user population, the sample size of only 22 people seems too small to make statements for the general population. In addition, all of the participants were recruited through a local newspaper, increasing the risk of a too homogeneous sample. In the second group, the sample consisted of only psychology students, mainly women (28 out of 33) and from a small age range (from 20 to 40), making the sample even more homogeneous, which even the authors themselves had to admit. This brings up the question of whether or not the the questionnaire can really be used on a wide population and not only by those people represented by the samples used in the validation studies. Further validation studies might therefore already be necessary based on the quality of sample subjects alone. In comparison, the samples used in the VisAWI's construction and validation process were larger and more representative of the general population in terms of both gender distribution as well as age range. Although recruited over the Internet, concerns that may come up through that, might be weakened when considering that the area of application for the VisAWI would be similar to the area of recruitment for the study participants. It also needs to be mentioned at this point, that sample sizes below 200 can be problematic for exploratory factor analysis processes (Howard, 2016). For the AttrakDiff2, this would mean that even if exploratory factor analysis would have been performed, results might have been prone to error. Sample sizes used in factor analysis for the VisAWI are above that mark of 200 people. For the AttrakDiff, one independent study was found that validated a newly created French version of the questionnaire (Lallemand, Koenig, Gronier, & Martin, 2015). Results were satisfying but further detail of the study can unfortunately not be discussed as part of this thesis due to the fact that the whole paper was written in French and understanding of it would therefore be limited. Besides this though, no further studies attempting to replicate neither the VisAWI nor the AttrakDiff could be found besides from the paper on the process of creating the VisAWI-S (Moshagen & Thielsch, 2013) which does not satisfy the need for independence either. When looking at the overall process of construction for the two questionnaires, it should be noted that both of them seem to follow a sound and scientific way, at least when looking at the steps performed on their own. Due to the concerns mentioned above, and also because of the overall effort put into construction and validation, it must be noted though, that the overall quality of the VisAWI seems to be higher than that of the AttrakDiff. Still, the AttrakDiff seems to be used more often than the VisAWI, when considering not only its citations but the examples from research that were provided above.

Implications for UX Research

While the VisAWI seems to represent adequate qualities for use in future scientific research, the AttrakDiff seems to lack some of them, mainly by bringing up questions about its validity on several occasions. Overall, this does not have to do with the quality of the construction methods used, but rather with the quality and amount of studies used to create and validate the AttrakDiff when compared with the VisAWI. It should be noted though, that both questionnaires present a very convenient way of application by providing an online platform that helps in eliminating sources of error stemming from the interaction of test subjects with any kind of researcher or lab assistant. These methods also provide a way of gathering more test subjects than with a pen and paper method, even if test subject numbers are unfortunately limited by the creators of the platform. Still, the AttrakDiff's validity is questionable and evidence for it seems to be lacking. What has to be considered though when using these questionnaires is, that they are old when comparing them with the progress in technology. When considering that the AttrakDiff2 is four years older than the first iPhone, questions of the actuality of its scales might come up. After all, computer technology and therefore user needs as well as user knowledge increase rapidly. Smart phone ownership steadily increases as well, such as in Switzerland where 72% of Internet users also used a mobile device for online access in 2017, showing a 29% increase since 2014 (Bundesamt für Statistik Schweiz, 2017b). In addition, emerging economies have also seen an increase in smart phone usage (Pew Research Center, 2016). Internet access therefore has probably seen a shift. Whether or not the questionnaires looked at in this thesis are still up to date in their measurements would probably have to be reevaluated, preferably by researchers independent from the original construction team. This seems to especially be necessary when considering, that some of the examples of use presented beforehand, used the questionnaires in handheld device (Choi & Lee, 2012; Grubert et al., 2015) and virtual reality settings (Gebhardt et al., 2014). Still, it can be said that the usage of any of these two questionnaires is better than the use of a self-made questionnaire that was only constructed for the purpose of one particular study, since efforts have been made to measure objectivity, reliability and validity. Transparency can also be seen as higher when validated questionnaires are used, especially when considering that items for self-made scales in UX were most often not provided to the reader (Bargas-Avila & Hornbæk, 2011). Further validation would though be recommended before using any of the two questionnaires, especially the AttrakDiff, to see whether or not they still are up to date with current technology.

Limitations and Future Research

First, the limited scope of this thesis made it only possible for the base papers of both the VisAWI and the AttrakDiff to be looked at. Even though the questionnaires themselves did not receive any updates after the papers presented here, at least to the extent of the knowledge provided by the literature research for this thesis, the underlying theories as well as the overall research of the authors did. Looking at the VisAWI-S might also bring up additional knowledge about the VisAWI itself and should be considered for future research. In addition, test construction is such a broad topic that not all methods of construction could be considered in this thesis. More aspects besides the base qualities of objectivity, reliability and validity as reported in the original papers could be investigated. Future research might for example look at examples from scientific research in a more structured manner to get a better grasp of the questionnaires' convergent and divergent validities. Looking at additional questionnaires used in the field of UX, such as the SUS (Brooke, 1996), and comparing their construction process with the AttrakDiff and VisAWI might also bring further knowledge about their quality in regards to possible alternatives. Finally, looking at recent developments in technology, such as the examples of handheld devices and virtual reality discussed in the research examples above, and comparing them with the questions provided for the two scales, might provide knowledge about their actuality and should be considered for future research together with revalidation of both the AttrakDiff and the VisAWI. The usage data hinted at by the creators of both the AttrakDiff as well as the VisAWI on their websites might also be of great value for better understanding of the two questionnaires and should hopefully be evaluated in future research.

Conclusion

Overall, it has to be said that both questionnaires were constructed in accordance with certain theoretical standards. While the VisAWI satisfies in terms of the major qualities of objectivity, reliability and validity, certain problems still remain with the AttrakDiff. Especially in regards to validity, satisfying evidence for the AttrakDiff is not present in the original papers. The two key issues for both questionnaires are lack of additional validation studies as well as the age of both questionnaires in comparison to rapid development in electronics. In addition, samples used for the AttrakDiff's construction and validation bear the risk of being too homogeneous and too small. Therefore, future research should aim at creating independent studies that revalidate the questionnaires and also try to find out, whether or not they still hold up under current technological environments, especially since both questionnaires are popular and offer convenient ways of application. To guarantee highest standards of scientific research, additional evaluation would have to be done on both questionnaires before further use.

Declaration of Scientific Integrity

The author hereby declares that he has read and fully adhered the Code for Good Practice in Research of the University of Basel.

References

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2689–2698).

Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4–7.

Bundesamt für Statistik Schweiz. (2017a). *Internetnutzung [Internet Usage].* Retrieved March 30, 2018, from `https://www.bfs.admin.ch/bfs/de/home/statistiken/kultur-medien-informationsgesellschaft-sport/informationsgesellschaft/gesamtindikatoren /haushalte-bevoelkerung/internetnutzung.html`

Bundesamt für Statistik Schweiz. (2017b). *Mobile Internetnutzung [Mobile Internet Usage].* Retrieved April 4, 2018, from `https://www.bfs.admin.ch/bfs/de/home/statistiken/kultur-medien-informationsgesellschaft-sport/informationsgesellschaft/gesamtindikatoren/haushalte-bevoelkerung/mobile-internetnutzung.html`

Cairns, P. (2007). HCI... not as it should be: inferential statistics in HCI research. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1* (pp. 195–201). British Computer Society.

Choi, J. H., & Lee, H. J. (2012). Facets of simplicity for the smartphone interface: A structural model. *International Journal of Human-Computer Studies*, *70*(2), 129–142.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297–334.

DeVellis, R. F. (2017). *Scale development: Theory and applications (Fourth Edition).* Thousand Oaks, CA: SAGE publications, Inc.

Diefenbach, S., Kolb, N., & Hassenzahl, M. (2014). The 'hedonic' in human-computer interaction: history, contributions, and future research directions. In *Proceedings of the 2014 conference on Designing interactive systems* (pp. 305–314). ACM.

Flavián, C., Guinalíu, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & management*, *43*(1), 1–14.

Gebhardt, S., Pick, S., Oster, T., Hentschel, B., & Kuhlen, T. (2014). An evaluation of a smart-phone-based menu system for immersive virtual environments. In *3D User Interfaces (3DUI), 2014 IEEE Symposium on* (pp. 31–34). IEEE.

Gnauk, B., Dannecker, L., & Hahmann, M. (2012). Leveraging gamification in demand dispatch systems. In *Proceedings of the 2012 joint EDBT/ICDT workshops* (pp. 103–110). ACM.

Grubert, J., Pahud, M., Grasset, R., Schmalstieg, D., & Seichter, H. (2015). The utility of magic lens interfaces on handheld devices for touristic map navigation. *Pervasive and Mobile Computing*, *18*, 88–103.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-computer interaction*, *19*(4), 319–349.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [AttrakDiff: A questionnaire for measuring perceived hedonic and pragmatic quality]. In *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187–196). Stuttgart: B. G. Teubner.

Hassenzahl, M., Diefenbach, S., & Göritz, A. (2010). Needs, affect, and interactive products–facets of user experience. *Interacting with computers*, *22*(5), 353–362.

Hassenzahl, M., Koller, F., & Burmester, M. (2008). Der User Experience (UX) auf der Spur: Zum Einsatz von www. attrakdiff. de [Tracking down the User Experience (UX): About the use of www.attrakdiff.de]. In *Tagungsband UP08* (pp. 78–82).

Stutthart: Fraunhofer Verlag.

Hassenzahl, M., Platz, A., Burmester, M., & Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 201–208).

Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & information technology*, *25*(2), 91–97.

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51–62.

Lallemand, C., Koenig, V., Gronier, G., & Martin, R. (2015). Création et validation d'une version française du questionnaire attrakdiff pour l'évaluation de l'expérience utilisateur des systèmes interactifs [Creation and Validation of a French Version of the AttrakDiff Questionnaire for the Evaluation of User Experience in Interactive Systems]. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *65*(5), 239–252.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group* (pp. 63–76). Springer, Berlin, Heidelberg.

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International journal of human-computer studies*, *60*(3), 269–298.

Law, E. L.-C., & van Schaik, P. (2010). Modelling user experience–an agenda for research and practice. *Interacting with computers*, *22*(5), 313–322.

Linares-Vásquez, M., Bavota, G., Cárdenas, C. E. B., Oliveto, R., Di Penta, M., & Poshyvanyk, D. (2015). Optimizing energy consumption of GUIs in Android apps: a multi-objective approach. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering* (pp. 143–154). ACM.

Lindgaard, G., Fernandes, G., Dudek, C., & Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & information technology*, *25*(2), 115–126.

MacDonald, C. M., & Atwood, M. E. (2013). Changing perspectives on evaluation in HCI: past, present, and future. In *Chi'13 extended abstracts on human factors in computing systems* (pp. 1969–1978).

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion (2. Auflage) [Test Theory and Questionnaire Construction (2. Edition)].* Berlin Heidelberg: Springer-Verlag.

Moshagen, M., & Thielsch, M. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies*, *68*(10), 689–709.

Moshagen, M., & Thielsch, M. (2013). A short version of the visual aesthetics of websites inventory. *Behaviour & Information Technology*, *32*(12), 1305–1311.

Pew Research Center. (2016). *Smartphone ownership and internet usage continues to climb in emerging economies.*

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.

Santayana, G. (1955). *The sense of beauty: Being the outline of aesthetic theory.* New York, NY: Dover Publications, Inc.

Solso, R. L. (2003). *The psychology of art and the evolution of the conscious brain.* Cambridge, MA: MIT press.

Stojmenovic, M., Pilgrim, C., & Lindgaard, G. (2014). Perceived and objective usability and visual appeal in a website domain with a less developed mental model. In *Proceedings of the 26th Australian computer-human interaction conference on designing futures: the future of design* (pp. 316–323). ACM.

Thanh Vi, C., Hornbæk, K., & Subramanian, S. (2017). Neuroanatomical correlates of perceived usability. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (pp. 519–532). ACM.

Thielsch, M. (2008). Ästhetik von Websites [Aesthetics of Websites]. *Wahrnehmung von Ästhetik und deren Beziehung zu Inhalt, Usability und Persönlichkeitsmerkmalen. Münster: MV Wissenschaft.*

Thielsch, M., & Moshagen, M. (2011). Erfassung visueller Ästhetik mit dem VisAWI. [Measuring Visual Aesthetics with the VisAWI]. In *Usability professionals* (pp. 260–265).

Thielsch, M., & Moshagen, M. (2014). *Visawi manual (visual aesthetics of websites inventory) and the short form visawi-s (short visual aesthetics of websites inventory).*

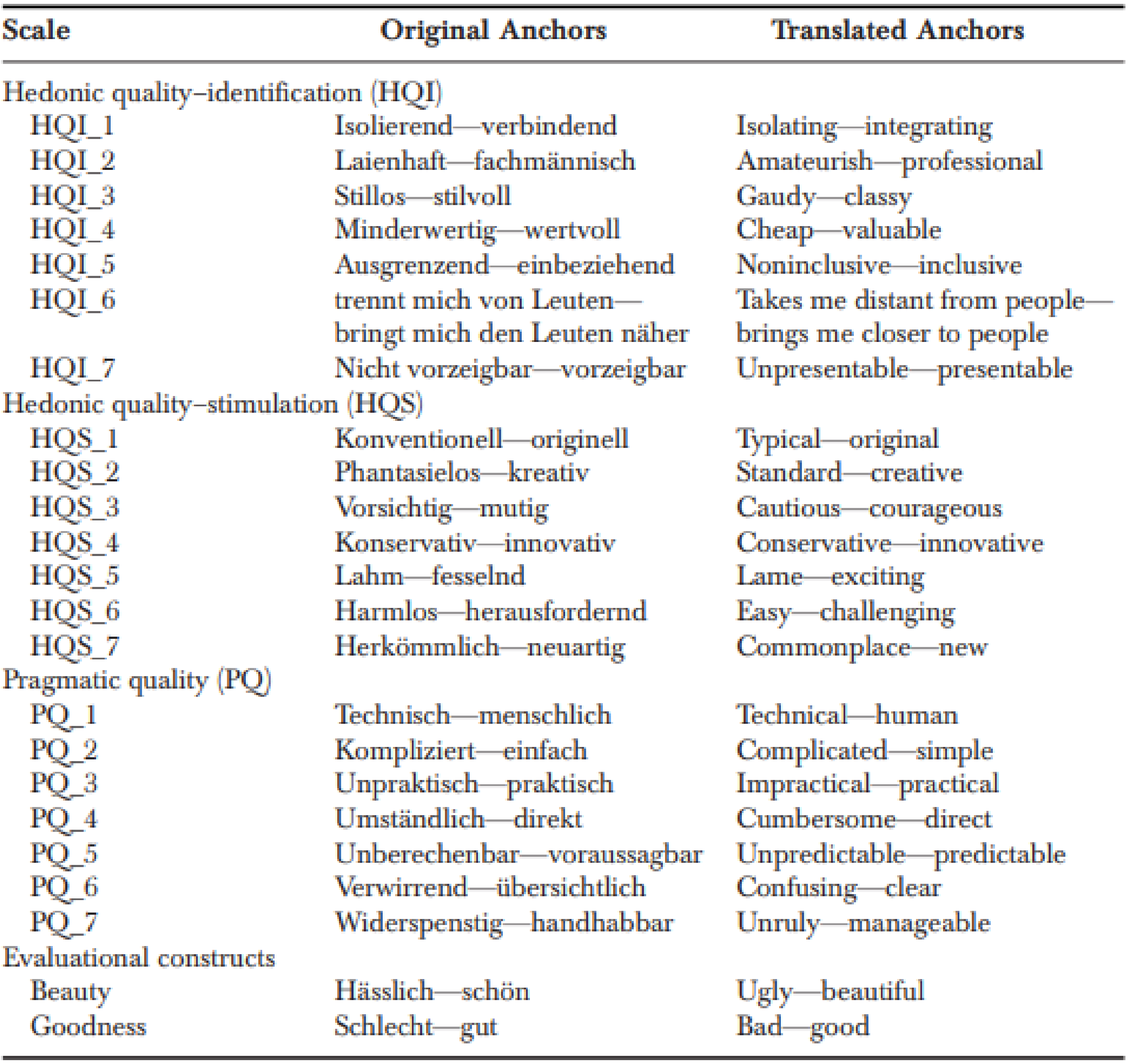
Tractinsky, N., Cokhavi, A., Kirschenbaum, M., & Sharfi, T. (2006). Evaluating the consistency of immediate aesthetic perceptions of web pages. *International journal of human-computer studies*, *64*(11), 1071–1083.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with computers*, *13*(2), 127–145.

Vermeeren, A., Law, E. L.-C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (pp. 521–530). ACM.

Appendix A

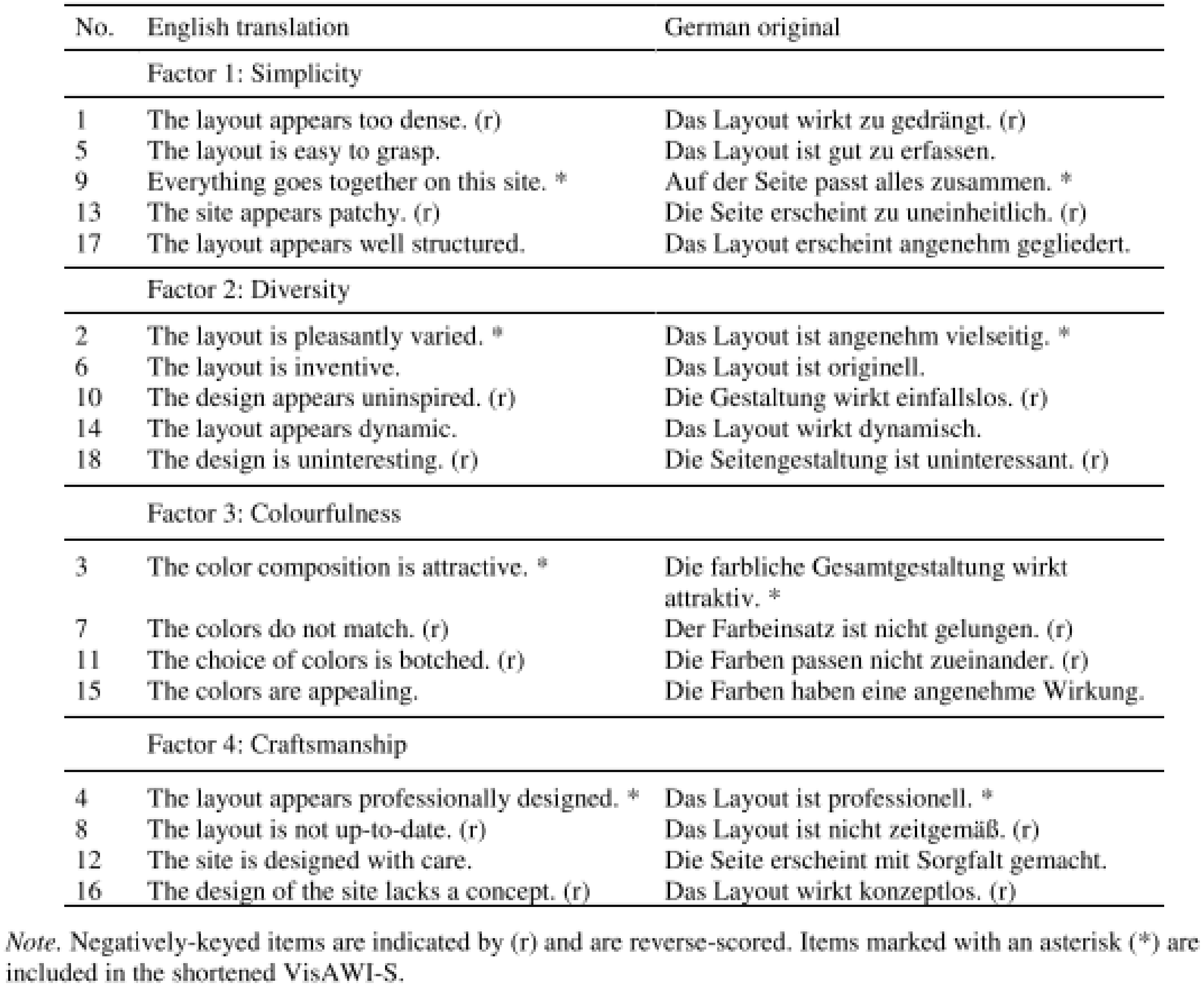
German and English items for the AttrakDiff2 as presented in Hassenzahl (2004, p.327).



Scale	Original Anchors	Translated Anchors
Hedonic quality–identification (HQI)		
HQI_1	Isolierend—verbindend	Isolating—integrating
HQI_2	Laienhaft—fachmännisch	Amateurish—professional
HQI_3	Stillos—stilvoll	Gaudy—classy
HQI_4	Minderwertig—wertvoll	Cheap—valuable
HQI_5	Ausgrenzend—einbeziehend	Noninclusive—inclusive
HQI_6	trennt mich von Leuten—bringt mich den Leuten näher	Takes me distant from people—brings me closer to people
HQI_7	Nicht vorzeigbar—vorzeigbar	Unpresentable—presentable
Hedonic quality–stimulation (HQS)		
HQS_1	Konventionell—originell	Typical—original
HQS_2	Phantasielos—kreativ	Standard—creative
HQS_3	Vorsichtig—mutig	Cautious—courageous
HQS_4	Konservativ—innovativ	Conservative—innovative
HQS_5	Lahm—fesselnd	Lame—exciting
HQS_6	Harmlos—herausfordernd	Easy—challenging
HQS_7	Herkömmlich—neuartig	Commonplace—new
Pragmatic quality (PQ)		
PQ_1	Technisch—menschlich	Technical—human
PQ_2	Kompliziert—einfach	Complicated—simple
PQ_3	Unpraktisch—praktisch	Impractical—practical
PQ_4	Umständlich—direkt	Cumbersome—direct
PQ_5	Unberechenbar—voraussagbar	Unpredictable—predictable
PQ_6	Verwirrend—übersichtlich	Confusing—clear
PQ_7	Widerspenstig—handhabbar	Unruly—manageable
Evaluational constructs		
Beauty	Hässlich—schön	Ugly—beautiful
Goodness	Schlecht—gut	Bad—good

Appendix B

German and English items for the VisAWI and the VisAWI-S as presented in Moshagen and Thielsch (2013, p.3).



No.	English translation	German original
	Factor 1: Simplicity	
1	The layout appears too dense. (r)	Das Layout wirkt zu gedrängt. (r)
5	The layout is easy to grasp.	Das Layout ist gut zu erfassen.
9	Everything goes together on this site. *	Auf der Seite passt alles zusammen. *
13	The site appears patchy. (r)	Die Seite erscheint zu uneinheitlich. (r)
17	The layout appears well structured.	Das Layout erscheint angenehm gegliedert.
	Factor 2: Diversity	
2	The layout is pleasantly varied. *	Das Layout ist angenehm vielseitig. *
6	The layout is inventive.	Das Layout ist originell.
10	The design appears uninspired. (r)	Die Gestaltung wirkt einfallslos. (r)
14	The layout appears dynamic.	Das Layout wirkt dynamisch.
18	The design is uninteresting. (r)	Die Seitengestaltung ist uninteressant. (r)
	Factor 3: Colourfulness	
3	The color composition is attractive. *	Die farbliche Gesamtgestaltung wirkt attraktiv. *
7	The colors do not match. (r)	Der Farbeinsatz ist nicht gelungen. (r)
11	The choice of colors is botched. (r)	Die Farben passen nicht zueinander. (r)
15	The colors are appealing.	Die Farben haben eine angenehme Wirkung.
	Factor 4: Craftsmanship	
4	The layout appears professionally designed. *	Das Layout ist professionell. *
8	The layout is not up-to-date. (r)	Das Layout ist nicht zeitgemäß. (r)
12	The site is designed with care.	Die Seite erscheint mit Sorgfalt gemacht.
16	The design of the site lacks a concept. (r)	Das Layout wirkt konzeptlos. (r)

Note. Negatively-keyed items are indicated by (r) and are reverse-scored. Items marked with an asterisk (*) are included in the shortened VisAWI-S.