INTERPRETING THE RESULTS FROM THE USER EXPERIENCE QUESTIONNAIRE (UEQ) USING IMPORTANCE-PERFORMANCE ANALYSIS (IPA)

Andreas Hinderks[1][a], Anna-Lena Meiners[2][b], Francisco José Domínguez Mayo[1][c], and Jörg Thomaschewski[2][d]

[1]*Department of Computer Languages and Systems, University of Seville, Seville, Spain*
[2]*University of Applied Sciences Emden/Leer, Emden, Germany*
andreas.hinderks@iwt2.org, meiners@ux-researchgroup.com, fjdominguez@us.es, joerg.thomaschewski@hs-emden-leer.de

Abstract: User Experience Questionnaire is a common and valid method to measure the User Experience (UX) for a product or service. In recent years, these questionnaires have established themselves to measure various aspects of UX. In addition to the questionnaire, an evaluation tool is usually offered so that the results of a study can be evaluated in the light of the questionnaire. As a rule, the evaluation consists of preparing the data and comparing it with a benchmark. Often this interpretation of the data is not sufficient as it only evaluates the current User Experience. However, it is desirable to determine exactly where there is a need for action. In our article we present an approach that evaluates the results from the User Experience Questionnaire (UEQ) using the importance-performance analysis (IPA). The aim is to create another possibility to interpret the results of the UEQ and to derive recommendations for action from them. In a first study with 219 participants, we validated the approach presented with YouTube and WhatsApp. The results show that the IPA provides additional insights from which further recommendations for action can be derived.

1 INTRODUCTION

In many companies, questionnaires are used to measure and evaluate the user experience of products and services, because UX questionnaires are a common quantitative way to measure of user experience (Lazar et al., 2010). There are numerous UX questionnaires in the literature, such as the Visual Aesthetics of Websites Inventory (VisAWI) (Moshagen and Thielsch, 2010), Standardized User Experience Percentile Rank Questionnaire (SUPR-Q) (Sauro, 2015) or the User Experience Questionnaire (UEQ) (Laugwitz et al., 2008). One aim of using a UX questionnaire is the request to derive recommendations for development in order to improve the product.

A well-known definition of user experience is given in ISO 9241-210 (ISO9241-210, 2010). Here user experience is defined as "a person's perceptions and responses that result from the use or anticipated use of a product, system or service". Thus, user experience is seen as a holistic concept that includes all types of emotional, cognitive or physical reactions concerning the concrete or even only the assumed usage of a product formed before, during and after use. However, the standard does not provide a definite list of factors or methods to measure user experience.

A different interpretation is to define user experience as a set of distinct quality criteria (Preece et al., 2015) that includes classical usability criteria or pragmatic qualities such as efficiency, controllability or learnability; and non-goal directed or hedonic quality criteria (Hassenzahl, 2001) such as

[a] https://orcid.org/0000-0003-3456-9273
[b] https://orcid.org/0000-0002-9803-1555
[c] https://orcid.org/0000-0003-3502-8858
[d] https://orcid.org/0000-0001-6364-5808

stimulation, fun-of-use, novelty, emotions (Norman, 2007), or aesthetics (Tractinsky, 1997). This has the advantage that it splits the general notion of user experience into a number of simple quality criteria, which describe distinct and relatively well-defined aspects of user experience that can be measured independently.

Questionnaires that measure the user experience take into account this complexity, since they usually compute values on different UX scales. A scale corresponds to a content-delimited quality characteristic of user experience, e.g. efficiency or originality. Depending on the questionnaire, different combinations of quality characteristics are measured. Standardized questionnaires are not a more or less random or subjective collection of questions, but result from a careful construction process. This process guarantees accurate measuring of the intended UX qualities. But on the other hand, a standard UX questionnaire is unable to measure user experience holistically (Osgood et al., 1978). A standardized questionnaire accurately measures the UX scales identified in the constructions such as stimulation, efficiency, attractiveness, etc.

The method presented in this paper is based on the User Experience Questionnaire (UEQ) (Laugwitz et al., 2008) and shows how to interpret the results from the UEQ by conducting an importance-performance analysis. We decided to use the UEQ because it is a well-known UX questionnaire and it is available in more then 20 languages. The objective of the UEQ is to allow a quick assessment done by end users covering a preferably comprehensive impression of user experience. It allows the users to express feelings, impressions, and attitudes that arise when experiencing the product under investigation in a very simple and immediate way. It consists of 26 items that are grouped into six scales (Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty). Each scale represents a distinct UX quality aspect.

The UEQ offers various options for interpreting the data. For example, the scales as well as the associated items can be interpreted individually. For each scale, there is also a benchmark that allows comparison with other data (Schrepp et al., 2017).

Another approach is the importance-performance analysis (IPA) (Martilla and James, 1977). An IPA measures customer satisfaction and presents it graphically so that recommendations for action can be made. Customer satisfaction is determined by querying the perceived importance and performance for a set of attributes. The result is displayed graphically in a plot. The recommendations for action are derived from the arrangement in the plot.

In this article, we present a method to interpret the results from the UEQ by conducting an importance-performance analysis (IPA).

Section 2 surveys the background and related work regarding the IPA. Section 3 outlines our method to interpret the results from the UEQ by conducting an IPA. Furthermore, we describe a first study to validate our method. In Section 4 we present the results of our study. Section 5 discusses the results of our study.

2 BACKGROUND AND RELATED WORK

As already described in the introduction, the importance-performance analysis (IPA) is one way of graphically representing the relationship between importance and performance for a set of attributes in a plot (Martilla and James, 1977).

There is no prescribed list of attributes for performing an IPA. The list of attributes must be determined during the concrete study (Martilla and James, 1977). In the literature, there are already proposals for selected products, for instance, Websites for airline companies (Öz, 2012) or Internet stores (Pokryshevskaya and Antipov, 2013). Another approach is to extract the items or scale from an existing questionnaire. Tontini (2016) has taken the items from the questionnaire e-SERVQUAL and used them as a set of attributes to evaluate online shopping sites. Also, there are various ways of creating a list of attributes.

The measurement of importance and performance is usually performed by directly putting the attributes on a seven-point rating scale, one item for importance, and one item for performance (Abalo et al., 2007; Azzopardi and Nash, 2013). There are other methods that derive importance indirectly from the results of performance (Bacon, 2003), for example, through multivariate regression analysis (Danaher and Mattsson, 1994) or a conjoint analysis (Danaher, 1997). This would have the advantage that only one item would have to be queried for importance and performance. The disadvantage, however, is the more reduced data quality (Bacon, 2003). In practice, direct measurement with two items per attribute has mostly established itself (Bacon, 2003).

The values from the items for each attribute are displayed in the IPA plot (Figure 1), where each attribute is assigned a point. The point is calculated by the value of the performance (x-axis) and importance (y-axis). The aim is to derive recommendations for action for each quadrant. The recommendation for action is derived from the relationship between importance and performance
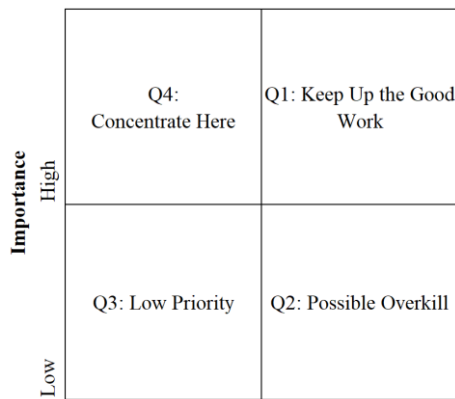
Figure 1: The Quadrants of the IPA Plot.

(Martilla and James, 1977). The underlying assumption is that a user is satisfied if his perceived importance is fulfilled. A measure of fulfilment is the value of performance.

The plot is typically divided into four quadrants (Figure 1):
- Q1: "Keep Up the Good Work"
- Q2: "Possible Overkill"
- Q3: "Low Priority"
- Q4: "Concentrate Here"

Figure 1 shows the four quadrants of the original IPA plot (Martilla and James, 1977). There are some illustrations in the literature where the axes are not in the same position. In this paper, we use the original usage of the axes of the IPA.

The first quadrant ("Keep Up the Good Work") represents great strengths and potential competitive advantages of a product or service. The user rates both the importance and the performance of the product equally highly. This means that there is no need for action for these attributes as they are balanced between importance and performance.

Attributes from quadrant 2 ("Possible Overkill") are rated relatively low by the user in the case of importance compared to performance. So importance is below performance, which means that the attributes are sufficiently developed. Further development of these attributes is, therefore, not necessary and would be inefficient since importance was more than fulfilled (Dwyer et al., 2012).

Attributes that fall under quadrant 3 ("Low Priority") are rated relatively low by the user both in terms of importance and performance. This means that no action is required for these attributes since both are balanced.

The fourth quadrant ("Concentrate Here") is the most important. Attributes from this quadrant are considered relatively important while performance is rated below average. These attributes offer the highest potential for perceptible improvement of the product. Further development of the product should, therefore, concentrate on these attributes.

3. RESEARCH METHODOLOGY

In this section, we will describe our approach in detail. The main idea behind our approach is to collect a dataset with the UEQ and then conduct an IPA with this dataset. In summary, we can use the results from the UEQ and we use the IPA to interpret the dataset. Our approach is divided into three different steps:
1. Step 1: Determine the attributes of the IPA.
2. Step 2: Selection of the questionnaire to gather the dataset for the IPA.
3. Step 3: First validation of the method from Step 2 by conducting a study with WhatsApp and YouTube.

The different steps are explained in more detail in the next three paragraphs.

3.1 Determine the attributes

There are no specifications as to how the attributes should be determined or selected (Section 2). Attributes should only represent quality criteria for the product (Martilla and James, 1977). For this reason, we have decided to use the UX scales of the UEQ as attributes for IPA.

For the IPA plot, data for the importance and performance for the particular set of attributes are required. The UEQ collects both the performance and importance. The performance is the actual value of the particular scale of the UEQ. The importance is additionally queried for each scale to calculate a UX KPI (Hinderks et al., 2019).

3.2 Selection of the questionnaire

The original UEQ consists of six UX scales Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty (Laugwitz et al., 2008). A modular extension of the 'User Experience Questionnaire' is the UEQ+ (Schrepp and Thomaschewski, in press). This new version of the UEQ (called UEQ+) has a modular structure so that the UX scales can be selected individually from a list for each test object. Step 1 is thus fulfilled.

In the first validation, we used both questionnaires, which are described in the next section.
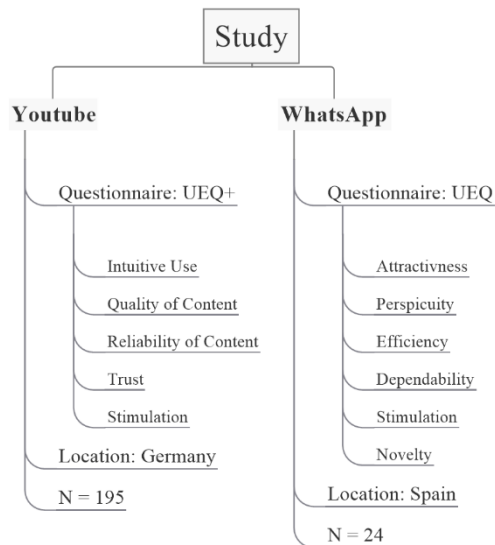
Figure 2: Overview of the Study.

3.3 First validation

The following study is intended to provide fundamental insights into our approach. We evaluated two products (YouTube and WhatsApp) with two different versions of the UEQ (UEQ+ and UEQ) (Figure 2).

For the UEQ+ we selected the following scales from the proposed list Intuitive Use, Quality of Content, Reliability of Content, Trust, and Stimulation. The two versions of the UEQ measure both performance and importance.

3.3.1 Object of the study

In this study, products with a high level of awareness were evaluated to ensure that the participants could assess the products. The test objects selected were YouTube and WhatsApp.

3.3.2 Purpose

The purpose of this study is to validate the use of IPA using the results from UEQ. The results should provide an understanding of the implementation of the IPA and the UEQ. It is to be determined whether the implementation of an IPA with the data of the UEQ provides good and interpretable results.

3.3.3 Quality focus

The main focus of the study is on validating the method by evaluating YouTube and WhatsApp. Here two specific aspects are emphasized. The choice is to focus on the confidence and scale consistency for every scale.

3.3.4 Context

The study was been conducted in Germany for YouTube and in Spain for WhatsApp through online and paper versions of the questionnaire. We collected the German dataset from the University of Applied Sciences Emden/Leer. The Spanish dataset was collected from the University of Seville.

A total of 219 participants took part in the study. In addition to the UEQ, we also asked for their age and gender. The participants assured us that they had used the product at least once a month.

The remaining answers were divided into 195 for YouTube and 24 for WhatsApp (Table 1).

Table 1: Number of Participants.

Test object	Total	
YouTube	195	(65 females, 123 males)
WhatsApp	24	(5 females, 18 males)
Total	219	

The average age is 32 years (31 for woman, 32 for men) for the German dataset and 23 years (22 for woman, 23 for men) for the Spanish dataset.

Table 2: Results from the UEQ for YouTube (Germany).

Scale	Performance			Importance		
	UEQ	Std.D.	Conf.	IMP	Std.D.	Conf.
Intuitive Use (INU)	1,736	1,388	0,195	0,933	1,639	0,230
Quality of Content (QOC)	1,249	1,377	0,193	1,267	1,509	0,212
Reliability of Content (ROC)	0,542	1,266	0,178	0,897	1,495	0,210
Trust (TRU)	-0,363	1,594	0,224	1,077	1,549	0,217
Stimulation (STI)	0,854	1,388	0,195	0,938	1,354	0,190

Table 3: Results from the UEQ for WhatsApp (Spain).

Scale	Performance			Importance		
	UEQ	Std.D.	Conf.	IMP	Std.D.	Conf.
Attractiveness (ATT)	1,722	0,553	0,221	1,875	1,154	0,462
Perspicuity (PER)	1,813	0,805	0,322	1,958	1,459	0,584
Efficiency (EFF)	1,635	0,516	0,206	2,375	1,096	0,438
Dependability (DEP)	1,267	0,651	0,260	2,083	0,881	0,352
Stimulation (STI)	1,000	0,699	0,280	1,375	1,583	0,633
Novelty (NOV)	0,281	1,025	0,410	1,292	1,429	0,572

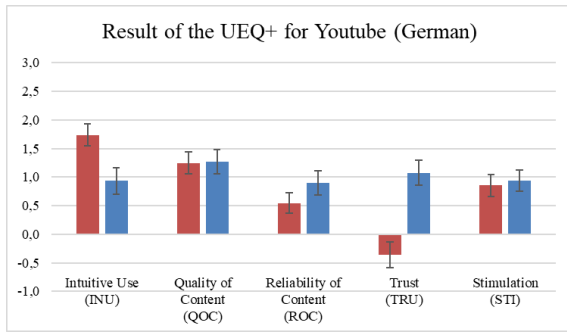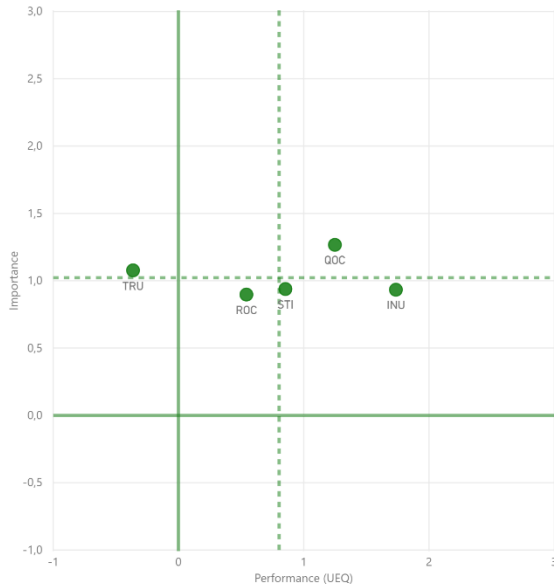Figure 3: Results from the UEQ+ for YouTube (Germany).

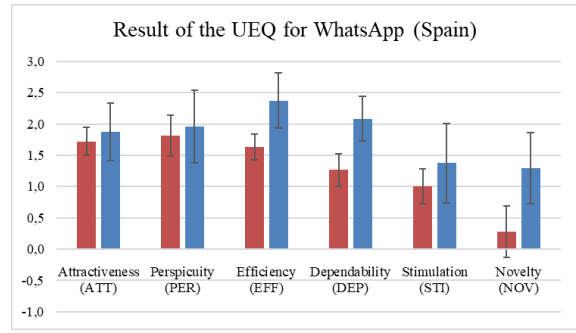Figure 5: Results of the UEQ for WhatsApp (Spain).

Figure 4: Results from the IPA for YouTube (Germany).

Figure 6: Results of the IPA for WhatsApp (Spain).

4. RESULTS

Thus, overall the participants had a slightly positive (> 1) or neutral (> -1 and < 1) impression concerning the user experience of YouTube (Table 2) and WhatsApp (Table 3). During the validation, we did not find any significant differences between men and women.

In Tables 1 and 2 the values for each scale are performance (UEQ value), and estimated importance, respectively. For each scale, the standard deviation and confidence are added. Figures 3 and 5 are the graphical interpretation of the values from Tables 1 and 2 The red bar (left) for each scale denotes performance and the blue bar (right) importance. The error bar represents confidence.

Reliability is typically estimated using the Standardized Cronbach Alpha coefficient (Nunnally and Bernstein, 2010). The Cronbach Alpha is a measure of the internal consistency of a questionnaire dimension (Cronbach, 1951). An analysis of the Cronbach Alpha coefficient showed that the single scales showed high consistency values for YouTube (INU: 0.93, QOC: 0.81, ROC: 0.89, TRU: 0.91, STI: 0.84). This is an indicator that the scales are sufficiently consistent (Cronbach, 1951). For WhatsApp, the Cronbach Alpha coefficient showed high consistency values except Efficiency, Dependability, and Stimulation (ATT: 0.75, PER: 0.75, EFF: 0.35, DEP: 0.41, STI: 0.27, NOV: 0.74). Due to the small group of participants for WhatsApp, the result was as expected. There is no general rule about how large the value should be. In practice, however, a value of > 0.7 has proved to be sufficient (Landauer et al., 1983).

Our approach presented in Section 3 was used to conduct an IPA. Figures 4 and 6 show the IPA plot for YouTube and WhatsApp. Each point in the IPA plot represents a scale calculated from the values for performance and importance. The coordinate axes with the solid line have the coordinate origin in the scale centre (0,0). On the other hand, the dotted coordinate axes have their coordinate origin in the mean value of all displayed scales. The coordinate axes are necessary for the interpretation of the scales to form the corresponding quadrants. From the IPA plot, the scales can be assigned to the respective quadrant. The overview of the assignment is shown in Tables 4 and 5

Table 1: Assignment Scales to IPA Quadrants for YouTube.

Scale	Scale Centre (0,0)	Scale Centre Avg
INU	Q1: Keep Up the Good Work	Q2: Possible Overkill
QOC	Q1: Keep Up the Good Work	Q1: Keep Up the Good Work
ROC	Q1: Keep Up the Good Work	Q3: Low Priority
TRU	Q4: Concentrate Here	Q4: Concentrate Here
STI	Q1: Keep Up the Good Work	Q2: Possible Overkill

Table 2: Assignment Scales to IPA Quadrants for WhatsApp.

Scale	Scale Centre (0,0)	Scale Centre Avg
ATT	Q1: Keep Up the Good Work	Q1: Keep Up the Good Work
PER	Q1: Keep Up the Good Work	Q1: Keep Up the Good Work
EFF	Q1: Keep Up the Good Work	Q1: Keep Up the Good Work
DEP	Q1: Keep Up the Good Work	Q1: Keep Up the Good Work
STI	Q1: Keep Up the Good Work	Q3: Low Priority
NOV	Q1: Keep Up the Good Work	Q4: Concentrate Here

5. DISCUSSION

The idea behind the IPA is to assign the individual scales to four different quadrants. Each quadrant then provides a recommendation for action for the respective scale (Section 2). In practice, there are two established methods for defining the quadrants (Bacon, 2003).

Method 1: Differentiation by the coordinate origin at (0,0). (solid line in Figures 4 and 6)

Method 2: Differentiation by the coordinate origin in the mean value of all scale values. (dotted line in Figures 4 and 6)

According to Method 1, there is potential for improvement in the scale *Trust* for YouTube (Q4: Concentrate Here). All other scales have been classified on YouTube in such a way that there is no need for action (Q1: Keep Up the Good Work). For WhatsApp, there is no need for action on any scale (Q1: Keep Up the Good Work).

In our analysis, we determined that classification according to Method 1 is not optimally usable for our approach. Method 1 assumes that participants will give a neutral rating of 0 (in the value range -3 and 3). It has been shown that in practical use, a neutral rating is more likely to be above 0, as the UEQ benchmark shows (Schrepp et al., 2017). In this respect, the usability of Method 1 is limited.

When using Method 2, the scales *Intuitive of Use* and *Stimulation* on YouTube are exceeded (Q2: Possible Overkill). This means that there is no potential for improvement for these scales, as the expectations of the users are more than fulfilled. For the scales *Reliability of Content* at YouTube and *Stimulation* at WhatsApp, the scales are balanced so that there is no need for action (Q3: Low Priority). For these scales, the value for performance and importance are low. The same applies to the scales *Quality of Content* for YouTube and *Attractiveness*, *Perspicuity*, *Efficiency*, and *Dependability* for WhatsApp (Q1: Keep Up the Good Work). The only difference is that the performance and importance were relatively highly rated. After all, these scales are also balanced. The two scales *Trust* at YouTube and *Novelty* at WhatsApp were ranked relatively low in terms of importance compared to the performance (Q4: Concentrate Here). This means that the user feels that these two scales are important, but are currently not being satisfactorily met. As a recommendation for action, it can be recommended that these two factors have to be improved.

In summary, it can be pointed out that Method 2 can give accurate statements regarding options for action in connection with the UEQ.

5.1 Comparing UEQ analysis and IPA

The analyses by the UEQ do not offer any recommendations for action. However, it is a good idea to compare the values for performance and importance directly. If the importance is higher than the performance, this scale should be improved. If this approach is applied to our studies, the *Reliability*

of Content and *Trust* scale on YouTube should be improved. At WhatsApp, the scales *Efficiency*, *Dependability*, *Stimulation* and *Novelty* should be improved.

Comparing the results from Methods 1 and 2 with these results, there are differences, which can be traced back to the IPA method itself. IPA considers the results from the UEQ relative to each other. This means that it is not the absolute difference between performance and importance that is relevant, but the relative difference to each other.

5.2 Enhancement of our approach

The results from Section 4 suggest that the IPA can be used with the results from the UEQ. In principle, this approach should also work for other questionnaires, which contain several scales clearly separated from each other in content. However, the UX questionnaire must measure both performance and importance. Otherwise our approach with the questionnaire is not usable.

5.3 Limitations

The approach presented in this paper could be validated in a first study. Further studies with other products should confirm the validity. In the study, it could not be validated whether the derived recommendations for action are suitable for practical use. This should be verified in further studies.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented an approach that analyses results from the User Experience Questionnaire (UEQ) using the importance-performance analysis (IPA). Our approach assigns the different scales of the UEQ to four different quadrants of the IPA plot. Each quadrant is assigned to a recommended course of action: Q1: 'Keep Up the Good Work', Q2: 'Possible Overkill', Q3: 'Low Priority', Q4: 'Concentrate Here'. We were able to validate this method in an initial study, in two countries, with a total of 219 participants, by evaluating YouTube and WhatsApp.

Our approach offers, in addition to the UEQ, another possibility to interpret the results of the UEQ. This can be useful for practical purposes and provides additional support for UEQ users.

Further research could examine whether our approach can be implemented in an organization.

However, it is necessary to validate our approach and implement it in a company in a real situation. Interpretability and acceptance should be emphasized. Also, it could determine whether our approach meets all requirements for practical usage.

REFERENCES

Abalo, J., Varela, J., and Manzano, V. 2007. Importance values for Importance–Performance Analysis: A formula for spreading out values derived from preference rankings. *Journal of Business Research*, 60, 115–121.

Azzopardi, E., and Nash, R. 2013. A critical evaluation of importance–performance analysis. *Tourism Management*, 35, 222–233.

Bacon, D.R. 2003. A Comparison of Approaches to Importance-Performance Analysis. *International Journal of Market Research*, 45, 1–15.

Cronbach, L.J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

Danaher, P.J. 1997. Using conjoint analysis to determine the relative importance of service attributes measured in customer satisfaction surveys. *Journal of Retailing*, 73, 235–260.

Danaher, P.J., and Mattsson, J. 1994. Customer Satisfaction during the Service Delivery Process. *European Journal of Marketing*, 28, 5–16.

Dwyer, L., Cvelbar, L.K., Edwards, D., and Mihalic, T. 2012. Fashioning a destination tourism future: The case of Slovenia. *Tourism Management*, 33, 305–316.

Hassenzahl, M. 2001. The Effect of Perceived Hedonic Quality on Product Appealingness. *INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION Volume 13(4)*, 481–499.

Hinderks, A., Schrepp, M., Mayo, F.J.D., Escalona, M.J., and Thomaschewski, J. 2019. Developing a UX KPI based on the User Experience Questionnaire. *Computer Standards & Interfaces*. Volume 65, 38-44

ISO9241-210. 2010. Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems. ISO 9241-210:2010.

Landauer, T.K., Galotti, K.M., and Hartwell, S. 1983. Natural command names and initial learning: A study of text-editing terms. *Commun. ACM*, 26, 495–503.

Laugwitz, B., Held, T., and Schrepp, M. 2008. Construction and Evaluation of a User Experience Questionnaire. In Holzinger, A. (Ed.), *HCI and Usability for Education and Work*, Springer Berlin Heidelberg, Berlin, Heidelberg, Volume 5298. 63–76.

Lazar, J., Feng, J.H., and Hochheiser, H. 2010. *Research methods in human-computer interaction*, Wiley, Chichester, West Sussex, U.K.

Martilla, J.A., and James, J.C. 1977. Importance-Performance Analysis. *Journal of Marketing Management*, Volume 41, 77–79.

Moshagen, M., and Thielsch, M.T. 2010. Facets of visual aesthetics. *International journal of human-computer studies*, 68, 689–709.

Norman, D.A. 2007. *Emotional Design: Why We Love (or Hate) Everyday Things*, Basic Books, New York.

Nunnally, J.C., and Bernstein, I.H. 2010. *Psychometric theory*, 3rd ed., Tata McGraw Hill Education Private Ltd, New Delhi.

Osgood, C.E., Suci, G.J., and Tannenbaum, P.H. 1978. *The measurement of meaning*, University of Illinois Press, Urbana-Champaign.

Öz, M. 2012. A research to evaluate the airline companies' websites via a consumer oriented approach. *Afr. J. Bus. Manage.*, 6., 4880-4900

Pokryshevskaya, E., and Antipov, E. 2013. Importance-Performance Analysis for Internet Stores: A System Based on Publicly Available Panel Data. *SSRN Journal*.

Preece, J., Rogers, Y., and Sharp, H. 2015. *Interaction design: Beyond human-computer interaction*, 4th ed., Wiley, Chichester.

Sauro, J. 2015. SUPR-Q: A Comprehensive Measure of the Quality of the Website User Experience. *Journal of Usability Studies*, 68–86.

Schrepp, M., Hinderks, A., and Thomaschewski, J. 2017. Construction of a Benchmark for the User Experience Questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Inteligence*, 4, 40–44.

Schrepp, M., and Thomaschewski, J. in press. Eine modulare Erweiterung des User Experience Questionnaire: Hinweise zur Anwendung im praktischen Projekten. In Gesellschaft für Informatik (Ed.), *Mensch und Computer 2019*.

Tontini, G. 2016. Identifying opportunities for improvement in online shopping sites. *Journal of Retailing and Consumer Services*, 31, 228–238.

Tractinsky, N. 1997. Aesthetics and apparent usability. In Pemberton, S. (Ed.), *the SIGCHI conference*. 115–122.