Why we love or hate our cars: A qualitative approach to the development of a quantitative user experience survey

Leandro Miletto Tonetto [a, b, *], Pieter M.A. Desmet [c]

[a] Universidade do Vale do Rio dos Sinos (Unisinos), Avenida Luiz Manoel Gonzaga, 744, Porto Alegre, CEP 90470-280, Brazil
[b] Zooma – Consumer Experience, Av. Carlos Gomes, 1610/704, Porto Alegre, CEP 90480-002, Brazil
[c] Department of Industrial Design, Delft University of Technology, Landbergstraat 15, 2628 CE, Delft, The Netherlands

ARTICLE INFO

ABSTRACT

This paper presents a more ecologically valid way of developing theory-based item questionnaires for measuring user experience. In this novel approach, items were generated using natural and domain-specific language of the research population, what seems to have made the survey much more sensitive to real experiences than theory-based ones. The approach was applied in a survey that measured car experience. Ten in-depth interviews were conducted with drivers inside their cars. The resulting transcripts were analysed with the aim of capturing their natural utterances for expressing their car experience. This analysis resulted in 71 categories of answers. For each category, one sentence was selected to serve as a survey-item. In an online platform, 538 respondents answered the survey. Data reliability, tested with Cronbach alpha index, was 0.94, suggesting a survey with highly reliable results to measure drivers' appraisals of their cars.

© 2016 Elsevier Ltd and The Ergonomics Society. All rights reserved.

1. Introduction

Interest in measuring user experience has progressively increased along with a growing interest in experience-driven design. Initially, design researchers borrowed instruments developed in the social sciences (for reviews, see Laurans et al., 2009; Poels and Dewitte, 2006; Desmet et al., 2016) to undertake these measurements. Later, domain-dedicated instruments were developed. These instruments, such as scales that measure subtle and mixed user emotions (Desmet, 2003), affective responses to interactive products (Hassenzahl et al., 2010), or user experience over time (Karapanos et al., 2009), proved to be much more sensitive to the particular characteristics of the user experience itself. As a result, over the last decade, an overwhelming amount and diversity of user experience measurement instruments have become available. In response to this proliferation, several inventories have been organized and frameworks proposed with the aim of providing a comprehensive overview (e.g., Obrist et al., 2009; Vermeeren et al., 2010; ENGAGE, 2006; Höök, 2008).

Besides the attempt to provide an overview, various authors have also made an effort to increase awareness regarding the added value of applying experience measures in design processes (Väänänen-Vainio-Mattila et al., 2008; Law et al., 2014). These studies have revealed some pertinent theoretical, methodological, and practical issues for the measurement of user experience in the context of design processes. One recurrent issue is related to whether user experience is best measured with *qualitative* or *quantitative* methods. This question seems to refer back to the classical distinction between reductionism and holism at the same time that it appears to divide the user experience research community in two (Law, 2011). One of the main advantages of qualitative data, which tends to be rich and detailed, is that it can inform about the causes of certain experiences and offer relevant insights for both envisioning design opportunities and formulating design requirements. In contrast, one of the main advantages of quantitative data, which tends to be simple and precise, is that it can provide an objective basis for critical decisions on design and developmental issues. Such objectivity may come in handy when trying to attract investments and convince stakeholders about the effectiveness of design decisions. In addition, quantitative data can be helpful in testing and improving user experience theories, which, in their turn, may provide universal principles applicable to a multitude of design contexts. In practice, many researchers choose to combine qualitative and quantitative approaches as, for instance, when using emotion measurement in combination with open interviews (see Desmet and Schifferstein, 2012).

---

* Corresponding author. Universidade do Vale do Rio dos Sinos (Unisinos), Avenida Luiz Manoel Gonzaga, 744, Porto Alegre, Brazil.
E-mail addresses: ltonetto@gmail.com (L.M. Tonetto), P.M.A.Desmet@tudelft.nl (P.M.A. Desmet).

Within the wide variety of approaches to assess user experience, the traditional questionnaire is still currently the most often used in the context of design and design research, both for obtaining quantitative and qualitative data (Bargas-Avila and Hornbæk, 2011). Surveys are easy to develop and administer; they do not require sophisticated instruments or software; and they can be easily tailored to the research challenge at hand. Item generation is typically theory-based, requiring researchers to translate theoretical factors or variables into survey items. Two often reported concerns about using such theory-based questionnaires relate to their face and ecological validity, which again seem to refer back to the reductionism versus holism debate (Law et al., 2014). Design researchers tend to be sceptical about the degree to which surveys are able to uncover actual 'real' feelings and responses. In addition, they tend to question if theory-based surveys are sufficiently sensitive to the richness and variety of people's responses to stimuli in real-life settings. Acknowledging these two issues, we propose a novel way of developing theory-based item questionnaires for measuring user experience. The key difference from traditional questionnaires is that natural and domain-specific language of the research population is actually used in the generation of the items. Our proposition, therefore, is that both face and ecological validity may be increased with the use of questionnaires that are, not only based on underlying theoretical factors, but also formulated in a natural language sensitive to real experiences of real people in real usage contexts.

In the specific case of this paper, the attempt of applying a natural-language approach to item-development takes place within the context of a car experience survey. The reason for choosing car design is that vehicles usually evoke strong emotions and rich user experiences (Desmet et al., 2000; Kamp, 2012; Hiemstra-van Mastrigt et al., 2015; Franz et al., 2012). Moreover, stakeholders in the automotive industry typically require quantitative measures to justify and evaluate experience-driven design initiatives (Saucken et al., 2014). In the paper, we first briefly introduce the theoretical basis for the experience survey. Then, we present the three steps undertaken for developing the natural-language based questionnaire. Next, we report on the application of the survey and its results. Finally, in the Discussion Section, we bring the paper to an end, exploring some challenges and future research possibilities.

2. Evaluation model: appraisal theory

The theoretical basis for our approach to survey development is appraisal theory, one of the most commonly used theories (either implicitly or explicitly) for understanding emotional responses both in standard emotion research (Frijda, 1993) and in design research (Desmet and Hekkert, 2002). An appraisal is a sense-evaluation of the 'relational meaning' of a stimulus event, which determines the perceived relevance of the event to one's well-being (Frijda, 1986; Lazarus, 1991). Events that are appraised as contributing to one's well-being evoke pleasant emotions, and those that are appraised as threatening or harming one's well-being evoke unpleasant emotions. Because appraisals mediate events, user goals, and emotions, they may trigger insights regarding the relationships among these variables (Desmet and Hekkert, 2002). Experience questionnaires are usually developed on the basis of sets of distinct appraisal components, where each one relates to a particular 'relational issue' of a stimulus event (Roseman, 2001; Scherer, 2001). Reviewing a series of appraisal theories, Demir et al. (2009) selected seven main appraisal components that are relevant for user experience in human–product interactions. These components are presented in Table 1 along with their respective key relational issue.

Demir et al. (2009) argued that these seven appraisal components (Table 1) facilitate a systematic and fine-grained analysis of emotions. At the same time, the authors acknowledged that these components are too abstract to be useful for design purposes and, thus, advised that they should be operationalized according to the particular design domain of interest. In our natural-language approach to developing experience questionnaires, this operationalisation was mediated by domain-relevant interviews with real users rather than directly done by the researchers themselves (as it is the common practice). The next section reports on the three steps involved in this procedure in more details.

3. Questionnaire development

3.1. Step 1: item generation

The first step was to conduct in-depth interviews to operationalize the appraisal components presented in Table 1 within the context of car experience, using natural and domain-relevant language. Respondents were recruited by e-mail. In this mail, they were asked to indicate to what extent they agreed with the sentences "I love my car" and "I hate my car", using a five-point Likert scale (ranging from completely agree to completely disagree).

Ten respondents were selected from 50 e-mails sent: 5 females and 5 males; 18–54 years old; 3 undergraduates, 4 professionals, and 3 graduate students; all Brazilians. Five completely agreed that they loved their cars, and five completely agreed that they hated their cars. Even though the makeup of the two groups was different, as seen in Table 2, these respondents were selected to ensure that the subsequent interviews would generate emotional exclamations, that is, evoke naturalistic sentences from the drivers when expressing appraisals in relation to their cars.

An interview-guiding list based on the seven appraisal components was developed to be discussed with the selected respondents. The interviews were conducted individually and took between 45 and 75 min. Time and place were determined by the respondents. For reasons of ecological validity, the interviews were conducted in the respondents' car; first while they just sat in the driver's seat and, after, while they were driving. The interviews were conducted while the car was parked. During the driving, the interviewers refrained from speaking and only registered spontaneous verbalizations. Interviews were recorded in video and transcribed afterwards.

All transcribed material was analysed according to content analysis (see Neuendorf, 2002). In line with general practice in this field, interviews were analysed until data saturation was reached, indicating that the ten respondents from the qualitative stage of this research were able to provide clear reasons to their appraisals. Categories of answers were identified, mainly by means of similarity. That is, similar sentences mentioned by different respondents were grouped together. Content analysis was initially developed separately by two researchers with more than ten years of experience with the technique. All transcriptions were read and generated a first draft of the categories, which was discussed on a meeting between them. In a collaborative process, the two researchers developed the final outline of categories.

This procedure resulted in 71 categories of answers, each related to at least one of the seven appraisal components. From each category, one sentence was selected to serve as the basis for the questionnaire items. In this selection, the following five basic guidelines for formulating survey items were taken into consideration (see Hinkin, 1998): (1) the language used is simple, short and familiar to the group of respondents (e.g., "It's a dream car." Sentences were selected from the interviews, using everyday

Table 1
Seven appraisal components in user experience.

Appraisal component	Key relational issue
Motive consistency	Does this event help me in attaining my goals and motives?
Intrinsic pleasantness	Is this event pleasant or enjoyable?
Expectation confirmation	Does this event confirm my expectations?
Agency	Who (or what) is responsible for this event?
Standards conformance	Does this event match with my social norms and standards?
Coping potential	Can I handle or change the event?
Certainty	How certain am I about the event?

Table 2
Characteristics of the individuals interviewed to develop the item generation.

Interviewee	Characteristics
Lover 1	Female, 20 years old, undergraduate student
Lover 2	Female, 24 years old, professional
Lover 3	Male, 31 years old, graduate student
Lover 4	Male, 34 years old, graduate student
Lover 5	Male, 54 years old, professional
Hater 1	Female, 21 years old, undergraduate student
Hater 2	Female, 32 years old, graduate student
Hater 3	Female, 46 years old, professional
Hater 4	Male, 18 years old, undergraduate student
Hater 5	Male, 37 years old, professional

language.); (2) each item assesses a single issue, avoiding combinations of issues (e.g., "It makes unbearable noises."); (3) items do not represent more than one theoretical construct (e.g., "Its engine is quiet" is representative of the appraisal component intrinsic pleasantness.); (4) leading formulations (questions phrased in a way that suggest expected responses) are avoided (e.g., we have refrained from using sentences such as "It is the best model I could pay for"); (5) items that are expected to generate little variance are avoided (e.g., sentences revealing ideas with which users obviously totally agree were not included, such as "a great car should never fail".).

3.2. Step 2: assessment of the item pool and questionnaire development

The second step was to prepare the 71 sentences collected for serving as survey items. The sentences were kept in the original language (Brazilian Portuguese) and were adapted as little as possible to preserve the drivers' genuine way of expressing themselves. Nevertheless, some small changes were made. Firstly, slang was eliminated. Secondly, some sentences were adapted to make sure that they were understandable when taken out of the original context of the conversation. Thirdly, emotional words (e.g., afraid and happy) were eliminated as much as possible since the focus of the questionnaire was not on particular emotions but on appraisals that evoke emotions. Regarding this last change, however, it is important to highlight that, whenever the omission of emotional words compromised the meaning of the sentences, they were kept in the original form. This kind of exception can be illustrated, for instance, by contrasting two connotatively different sentences included in the item pool: "It has already given me so many problems that I'm afraid of it" and "This car has just given me problems". These two sentences would end up having very similar meanings, if the clause "that I'm afraid of it" had been omitted from the first sentence.

The questionnaire was built in an online platform. It started with general questions about the respondents' demographics and cars. Next, respondents indicated their level of agreement with each one of the 71 sentences, using 5-point Likert-type scales (1 = "completely disagree"; 5 = "completely agree"). Sentences were randomized among respondents to control for order and tiredness effects.

3.3. Step 3: content and face validation

The third step was to authenticate both the content and the face validity of the questionnaire. For this survey, content validity referred to the extent to which it sufficiently represented all seven appraisal dimensions (Table 1). Content validity is typically judged by experienced (and recognized) subject matter experts. Here, the judge was a leading expert in experience design research (board member of the International Design and Emotion Society, editor of peer reviewed design research journals, and professor of design aesthetics). In a meeting with the authors, the judge discussed the relevance and form of all items, suggesting several minor modifications. While content validity refers to what the survey actually measures, face validity refers to what the survey appears to measure in the eyes of the respondent population, and the adequacy of the language and item presentation. Face validity was determined with 61 undergraduate students from a southern Brazilian university. They first answered the survey in a media lab, and then reflected on each item. This step resulted in some other additional minor modifications of some items; the final set of items is presented in Table 4.

3.4. Application study

The final version of the natural-language questionnaire was tested in a study with 710 Brazilian respondents. Respondents were recruited by e-mail. The data of 538 respondents (75.8%) who completed the whole survey were included in the data analysis. Briefly, information about the respondents and their vehicles may be summarized as follows:

- mean age: 31.2 (SD = 10.94);
- gender: 56.7% female and 43.3% male;
- level of education: 30.1% undergraduate students, 22.7% professionals holding a bachelor degree or equivalent, and 38.5% holders of a master degree;
- car brands: 16.2% Chevrolet, 15.8% Volkswagen; 14.9% Fiat, and 14.3% Ford;
- car condition at purchase: 65.2% new and 34.8% used;
- car use frequency: 76.9% on a daily basis;
- car users: 51.7% single users and 24.5% main users, but not the only ones.

Respondents filled out the questionnaire individually, online, at a place and time of their convenience.

4. Results

Data reliability was tested with Cronbach's alpha. The result was

0.94, indicating an excellent internal consistency. The KMO measure of sampling adequacy was calculated to determine if the data was suitable for factor analysis. The result was 0.93, which is exceptionally high and suggested that the data was very suitable for factor analysis (Kaiser, 1974). In addition, we used Bartlett's test of sphericity that probes whether the correlation matrix is an identity matrix, in which case it would reveal that the factor model was inappropriate. The result, $16664.81 - p < 0.01$, however, indicated that the factorial model was adequate (Norusis, 1994). The communalities also presented a mean of 0.59, implying that each variable shared a good amount of variance in the analysis with the other measured variables. All items in the scale had communality equal or above 0.46.

While the scale items were developed according to the seven theoretical factors, the factorial extraction suggested 15 factors, with the lowest one explaining 1.72 of the total variance. When results suggest more than ten factors, the eigenvalues are usually below 1.0, which would indicate that we should not accept this number of factors. This was not the case in this survey, since all eigenvalues were above 1.0. It was found that the 15 factors explained 58.989% of the total variance from the group of variables that were part of the instrument. Even though the variance explained by each factor was relatively balanced among them, the distance between the first (23.32) and the second (7.19) factors was higher. The Varimax orthogonal rotation showed a better distribution of the variance among factors. All these results can be seen in Table 3.

4.1. Factors composition

Item distribution in factors, considering their factorial extraction observed after the Varimax rotation and the Kaiser normalisation, is shown in Table 4.

Even though factors from 11 to 15 are composed by one item each, we decided to keep them in the survey, due to their communality scores. Observing their contents, it is not surprising that they did not charge in any other factor, since they do not seem to share a lot in meaning with any other general factor. Some of these last 5 factors even had some of the highest communality scores in the 71-item survey.

Each one of the 15 factors was named according to its item contents. A brief description of each factor's content can be found in Table 5.

Motive consistency seems to be related to almost all of the 15 factors in Table 5. This finding is compatible with appraisal theory, which proposes that motive consistency is the 'primary appraisal' that determines if an emotion is experienced and if this emotion is positive or negative. The other appraisal components are secondary and determine what particular emotion is experienced (Roseman, 1991). Nonetheless, the other six appraisals are also represented by one or more factors: intrinsic pleasantness (factor 3); expectation confirmation (factor 13); agency (factor 14); standards conformance (factor 10); coping potential (factors 1 and 14); certainty (factor 11).

4.2. Convergent validity

Convergent validity, in this case, is the degree to which a factor converges on the global scale results. It shows that assessment is related to what it should be related to. To test for convergent validity, a multiple linear regression analysis was conducted, with the item's global mean score as a dependent variable and the mean score for all items in each factor as an independent variable. This analysis revealed a significant Adjusted R Square (0.998), indicating that most of the variance of the global results is explained by the 15 factors that are part of the scale. Table 6 shows how much each factor predicts the global scale results.

It can be observed that the highest results are 0.404 (factor 2) and 0.338 (factor 1), both of them considerably distant from all the others, which means that the first two factors themselves can predict most of the global results of this research.

5. Discussion and conclusion

In an inventory of user experience research, Bargas-Avila and Hornbæk (2011) found that half of the publications that used questionnaires employed self-developed instruments without providing readers with information about which items were used or on what basis these items were formulated. These authors claimed that using self-made items without such relevant information is unscientific, preventing readers from understanding and assessing the quality of the research approach and its outcomes. Their observation highlights an issue that is usually undervalued in the debate about the usefulness of questionnaires in experience-driven design: how survey items should be developed to maximise their contribution. Most of the attention seems to be placed on the contribution of quantitative research only.

The research reported in this paper aimed at taking a step towards exploring this issue, putting forward a novel way of developing user experience surveys. More specifically, it argued that the

Table 3
Factor analysis results before and after the rotation.

Factors	Initial extraction			After rotation		
	Eigenvalues	% Variance	% Cumulative variance	Eigenvalues	% Variance	% Cumulative variance
1	16.555	23.318	23.318	8.356	11.769	11.769
2	5.101	7.185	30.502	7.957	11.207	22.976
3	3.650	5.141	35.643	4.975	7.007	29.983
4	2.091	2.945	38.588	2.227	3.137	33.120
5	1.948	2.744	41.332	2.167	3.052	36.173
6	1.674	2.358	43.690	2.114	2.977	39.150
7	1.432	2.017	45.707	2.048	2.884	42.034
8	1.399	1.971	47.678	1.959	2.759	44.793
9	1.300	1.832	49.509	1.707	2.405	47.198
10	1.280	1.803	51.312	1.595	2.247	49.445
11	1.187	1.672	52.985	1.465	2.063	51.508
12	1.112	1.566	54.551	1.416	1.994	53.502
13	1.079	1.519	56.070	1.350	1.901	55.403
14	1.048	1.476	57.546	1.325	1.867	57.269
15	1.025	1.443	58.989	1.221	1.720	58.989

Table 4
Factors, their respective items and the rotated component matrix.

Factor	Factorial charge	Item	Item description
1	0.790	66	It has already given me so many problems that I'm afraid of it.
	0.746	43	This car has just given me problems.
	0.739	65	I go through situations with my car, caused by it, that make me angry.
	0.717	71	My car is a Pandora's box. Anything can happen to it.
	0.705	47	I am traumatised by this car.
	0.703	70	I need to be alert to drive my car, due to the problems I have with it.
	0.660	67	I don't feel safe in my car, because it fails and I don't know the reason.
	0.585	35	I thought it was going to give fewer problems than it has.
	0.579	63	The impression I have is that something will get broken in the car.
	0.550	33	It never lets me down.
	0.526	40	I have confidence in it, because it has never caused any problems.
	0.521	29	It makes unbearable noises.
	0.517	68	My car makes me embarrassed sometimes.
	0.483	59	I believe it's not going to give many problems, I think it's going to be very durable.
	0.479	56	It doesn't give me any surprises, like things that fail or fail slowly.
	0.426	62	I have confidence in the car, that if something goes wrong, I can get good advice.
	0.423	38	It obeys the orders as I want it to.
	0.269	18	I don't feel good when I look at it.
2	0.766	2	It seems like it's a part of me.
	0.735	4	I identify myself with my car, since I have had many happy times with it.
	0.734	1	My car suits me.
	0.704	6	My car says something about me.
	0.688	10	It gives me a feeling of freedom, more than other cars.
	0.687	7	It's able to blend itself into my way of life.
	0.657	26	The other cars are not as fun as mine.
	0.573	8	I'm a fan of my car; it allows me to do everything I want to do with it.
	0.552	42	I feel more confident to drive it than I felt before with other cars.
	0.546	17	I like everything about it.
	0.536	46	There's no better car than this.
	0.527	36	It seems that I started to drive better, after I got this car.
	0.511	5	It's perfect for me, there are no problems.
	0.511	9	It's a dream car.
	0.449	37	It's a car that I have adapted myself very well to.
	0.412	34	It represents everything that I have expected from it.
	0.403	49	It was more valuable to buy this car, than any other car in its category.
	0.377	16	There are several advantages that have convinced me to buy the car.
	0.272	25	I see other cars and I don't feel jealous of them.
3	0.657	21	Its engine is quiet.
	0.634	22	It doesn't have internal noises.
	0.606	20	My car is comfortable.
	0.538	50	It has everything I need, it's really complete.
	0.533	28	Its internal parts are good to touch.
	0.496	64	Even in adverse situations, like traffic jams, I like to be in the car, due to the comfort it offers.
	0.493	23	I like its internal space.
	0.488	19	It smells/smelled new, really good.
4	0.758	41	There is something in my car's structure that seems fragile.
	0.741	55	I feel that it's fragile, because it has many plastic components.
	0.363	14	My car is more resistant than others.
5	0.566	39	I don't have any technical difficulties, everything is self explanatory.
	0.382	52	Its performance is better than other cars in the same category.
	0.374	27	It's really easy to adjust my car to my needs.
	0.311	3	I get to do all I want to with my car.
	0.251	11	It's easy to park.
6	0.693	30	People show respect for this car because it's big.
	0.556	31	My car is robust.
7	0.729	61	Its maintenance is expensive.
	0.679	58	I don't believe the maintenance will be too expensive.
	0.530	60	There is a shortage of parts, when I need to find them.
8	0.757	12	One of the reasons I bought this car is that it consumes almost no fuel.
	0.753	51	It consumes very little fuel, compared to other cars in its category.
9	0.765	45	Its insurance isn't so expensive.
	0.709	44	The annual road tax isn't so expensive.
	0.521	13	It isn't a target for thieves.
10	0.665	54	It doesn't have the same things that other cars have in its category.
	0.454	32	Its design is out-dated.
	0.398	53	I don't think this car is as comfortable as others I have seen.
11	0.507	57	I know what's going to happen to the car in the future; I won't have any surprises.
12	0.586	15	I believe it should be easy to resell my car, if I wanted to.
13	0.654	48	For the price paid, it should have been better made.
14	0.800	69	The biggest problem of this car is the driver.
15	0.794	24	It does not have areas that you can not see.

way in which items are generated could influence the internal consistency and the ecological validity of this kind of instrument.

Table 5
Factor topics.

Factor	Factor topic
1	Lack of reliability and safety
2	Personal identification
3	Comfort (due to lack of sensorial displeasure)
4	Fragility
5	Ease for use
6	Social power
7	Maintenance costs
8	Fuel costs
9	Taxes and insurance costs
10	Inferiority to other cars
11	Predictability in use
12	Reselling potential
13	Lack of value for money
14	Problems caused by the driver
15	Visual breadth

Table 6
Regression analysis.

Factors	Unstandardized coefficients	Standardized coefficients	T
(Constant)	0.010		1.316[a]
1	0.261	0.338	132.962[b]
2	0.273	0.404	145.279[b]
3	0.099	0.168	57.771[b]
4	0.040	0.069	32.533[b]
5	0.071	0.095	37.779[b]
6	0.028	0.060	27.657[b]
7	0.046	0.081	38.323[b]
8	0.026	0.056	28.209[b]
9	0.041	0.071	37.057[b]
10	0.044	0.075	35.986[b]
11	0.015	0.035	18.586[b]
12	0.014	0.029	14.727[b]
13	0.012	0.028	14.141[b]
14	0.014	0.033	18.201[b]
15	0.014	0.034	18.377[b]

[a] Significant differences at $\alpha = 0.05$.
[b] Significant differences at $\alpha = 0.001$.

Although it is common practice to formulate survey items in a simple and direct language, appraisals towards products do not seem to be expressed in such straightforward terms in real life.

In the development of the car experience survey, it was observed that, when drivers verbally reported their experiences, they often connected attributes that are not usually combined in traditional survey items. By respecting these natural combinations in the survey development, the research seemed to have gained in ecological validity. For example, the sentence "I have confidence in it, because it has never caused any problems" states a causality relation. According to item-development guidelines, it should have been split into two items. However, respondents understood confidence only in relation to this particular cause. Thus, working with combined statements represents a much more real-life appraisal in relation to this particular product.

Therefore, the main contribution of this research was to demonstrate that, when using a natural and domain-specific approach to develop user experience surveys, researchers are more likely to end up with consistent research instruments. To integrate this naturalistic — qualitative dimension to a survey, researchers need to be flexible regarding the traditional ways of developing survey items. On balance, we believe that this research may represent an advance in the design and emotion field. It not only moves towards holism in evaluating real-life experiences, as advocated by Law (2011), but it also provides a strong basis to measure experience, which is still a key issue in contemporary user

research.

It is also worth pointing out a secondary contribution, on a theoretical level: a better understanding of the role of appraisal theory in design studies. As discussed before, there was not a complete convergence between theoretical factors (appraisal components) and those extracted from the research data. For example, motive consistency was an appraisal component that seemed to be related to most of the 15 extracted factors. This outcome seems to indicate that a one-to-one correspondence between real-life appraisals to products and specific theoretical appraisal components or single questions about them is unreal. Theoretical constructs have consistently been useful in psychological research, among other functions, to inspire the design of data collection instruments. They may not reveal what happens in real life, but still be a source of input to collect and analyze data. Thus, it is normal and expected, e.g., that users buy cars that are more aesthetically appealing and pleasant to their senses. In Appraisal Theory, it would be both interpreted as motive consistency and intrinsic pleasantness.

The survey investigates appraisals that tend to evoke positive experiences, which can be useful to the industry. Some of these appraisals can be very factual, providing a lower level of insights to design better cars (e.g., "Its engine is quiet"). On the other hand, several other appraisals are less objective and allow researchers to foresee ways in which car design may be able to evoke better experiences than the current ones (e.g., "It seems like it's a part of me" and "It gives me a feeling of freedom, more than other cars"). Therefore, practical and subjective appraisals merge and determine the overall quality of user experience. Ignoring the non-factual appraisals is not recommended. As regression analysis indicates, factor 2 (personal identification), which is a very subjective factor, explains most of the variance of the global results of the survey.

Even though the results seem to support the idea that our approach allows for a consistent survey development, some challenges and limitations must be pointed out. One challenge is the amount of time spent in the elaboration of the survey. A traditional questionnaire can be developed in a few hours, while developing a survey in the lines of the proposed approach can take several weeks, since it requires interviewing users, analysing the content of their verbalizations, and generating the survey items based on qualitative analysis. Thus, general theory-based surveys continue to be more convenient when researchers have limited time.

In relation to the specificity and the depth of the analysis, a primary limitation is that this kind of research instrument is specific to evaluate user experience with a single product (category), which means that all survey items will be strictly related to one category. This limited scope, however, seems to be a trend in psychometrics: the development of specific surveys, instead of general ones. Although questionnaires gain in specificity, data reliability and validity, they lose in terms of development agility and availability, as well as in terms of the "purity" of traditional academic procedures.

Another limitation is that the approach is not suitable to develop a single survey version to be used with extremely distinct research populations. Considering that the actual products evaluated differ from culture to culture, it is clear that appraisals to them may also change, as well as the emotional outcomes from user—car interaction due to these appraisals. In addition, assuming that people in different cultures express appraisals in different ways, it also seems inadequate to develop research instruments to be used in cross-cultural studies. On the other hand, the industry is certainly more benefited from the use of a highly consistent instrument that reveals real-life appraisals than a general one, since the specificity of the product-related contents and the adequacy of the appraisals can deliver results to designers that allow for safe decision-making.

The type of relationship established between users and cars may also be different. In Brazil, where the data collection was developed, recently the government lowered taxes to stimulate citizens to purchase new cars. In addition, the public transportation system is known by being chaotic in most major cities, which can also stimulate people to have closer relationships with their cars due to intense use. Extremely different user-car relationships may be observed in distinct cultures, such as Denmark and The Netherlands, where citizens are stimulated to use other types of transport.

An important limitation of this study is that participants were not asked how recently they have been in their cars. It is known that the experience of use may change through time, and this variable should be addressed in further studies.

While this research is an initial effort in shaping a more naturalistic way of developing user experience questionnaires, based on respondents' natural and domain-specific language, further studies are required to make it even more solid. A research agenda on the topic could include the development of car experience surveys with the same approach but in different cultures, as well as translations of the current survey to other languages in order to compare results with the original findings in this paper. It is also important to develop further research instruments to evaluate users' appraisals of other categories of products, using the same approach, to access whether their results would also be highly consistent or not. Lastly, considering both the needs of the industry, to test products worldwide, and those of the academia, to have proper instruments and to accumulate knowledge based on similar scientific criteria, some effort could be done to try to develop reality-driven instruments suitable to cross-cultural studies.

Acknowledgements

References

Bargas-Avila, J.A., Hornbæk, K., 2011, May. 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. ACM, New York, pp. 2689–2698.

Demir, E., Desmet, P.M.A., Hekkert, P., 2009. Appraisal patterns of emotions in user–product interaction. Int. J. Des. 3 (2), 41–51.

Desmet, P.M.A., 2003. Measuring emotion: development and application of an instrument to measure emotional responses to products. In: Blythe, M.A., Monk, A.F., Overbeeke, K., Wright, P.C. (Eds.), Funology: from Usability to Enjoyment. Kluwer Academic Publishers, Dordrecht, pp. 111–123.

Desmet, P.M.A., Hekkert, P., 2002. The basis of product emotions. In: Green, W., Jordan, P. (Eds.), Pleasure with Products, beyond Usability. Taylor & Francis, London, pp. 60–68.

Desmet, P.M.A., Hekkert, P., Jacobs, J.J., 2000. When a car makes you smile: development and application of an instrument to measure product emotions. Adv. Consum. Res. 27, 111–117.

Desmet, P.M.A., Schifferstein, N.J.H., 2012. Emotion research as input for product design. In: Beckley, J., Paredes, D., Lopetcharat, K. (Eds.), Product Innovation Toolbox: a Field Guide to Consumer Understanding and Research. John Wiley & Sons, Hoboken, NJ, pp. 149–175.

Desmet, P.M.A., Vastenburg, M.H., Romero, N., 2016. Mood measurement with pick-A-mood: review of current methods and design of a pictorial self-report scale. J. Design Res. (in press).

ENGAGE, 2006. Report on the Evaluation of Generative Tools and Methods for 'Emotional Design'. Deliverable D15.3. EU Project Engage 520998.

Franz, M., Durt, A., Zenk, R., Desmet, P.M.A., 2012. Comfort effects of a new car headrest with neck support. Appl. Ergon. 43 (2), 336–343.

Frijda, N.H., 1986. The Emotions. Cambridge University Press, Cambridge, UK.

Frijda, N.H., 1993. The place of appraisal in emotion. Cognit. Emot. 7 (3–4), 357–387.

Hassenzahl, M., Diefenbach, S., Göritz, A.S., 2010. Needs, affect, interactive products – facets of user experience. Interact. Comput. 22, 353–362.

Hiemstra-van Mastrigt, S., Kamp, I., van Veen, S.A.T., Vink, P., Bosch, T., 2015. The influence of active seating on car passengers' perceived comfort and activity levels. Appl. Ergon. 47, 211–219.

Hinkin, T., 1998. A brief tutorial on the development of measures for use in survey questionnaires. Organ. Res. Methods 1 (1), 104–121.

Höök, K., 2008. Human-machine Interaction Network on Emotions (HUMAINE). Final Report on WP9. Stockholm University, Stockholm.

Kaiser, H., 1974. An index of factorial simplicity. Psychometrika 39, 31–36.

Kamp, I., 2012. The influence of car-seat design on its character experience. Appl. Ergon. 43 (2), 329–335.

Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J., 2009. User experience over time: an initial framework. In: Proceedings of CHI'09. ACM, New York, pp. 729–738.

Law, E.L.C., 2011. The measurability and predictability of user experience. In: Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems. ACM, New York, pp. 1–10.

Law, E.L.C., van Schaik, P., Roto, V., 2014. Attitudes towards user experience (UX) measurement. Int. J. Hum. Comput. Stud. 72 (6), 526–541.

Laurans, G.F.G., Desmet, P.M.A., Hekkert, P., 2009. Assessing emotion in interaction: some problems and a new approach. In: Guenand, A. (Ed.), Proceedings of the 4th International Conference on Designing Pleasurable Products and Interfaces, October 13–16, 2009, Compiegne, France.

Lazarus, R.S., 1991. Emotion and Adaptation. Oxford University Press, Oxford.

Neuendorf, K.A., 2002. The Content Analysis Guidebook, vol. 300. Sage Publications, Thousand Oaks.

Norusis, M., 1994. SPSS Professional Statistic 6,1. SPSS, Chicago.

Obrist, M., Roto, V., Väänänen-Vainio-Mattila, K., 2009. User experience evaluation: do you know which method to use?. In: Extended Abstracts of CHI'09. ACM, New York, pp. 2763–2766.

Poels, K., Dewitte, S., 2006. How to capture the heart? Reviewing 20 years of emotion measurement in advertising. J. Advert. Res. 46 (1), 18–37.

Roseman, I.J., 1991. Appraisal determinants of discrete emotions. Cognit. Emot. 5 (3), 161–200.

Roseman, I.J., 2001. A model of appraisal in the emotion system: integrating theory, research, and applications. In: Scherer, K., Schorr, A., Johnstone, T. (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, New York, pp. 68–91.

Saucken, C. v., Michailidou, I., Kremer, S., Lindemann, U., 2014. Motive-oriented design: helping automobile engineers to take the user's perspective. In: Proceedings of DUXU 2014, vol. 4. Springer, Amsterdam, pp. 370–377.

Scherer, K.R., 2001. Appraisals considered as a process of multilevel sequential checking. In: Scherer, K., Schorr, A., Johnstone, T. (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, New York, pp. 92–120.

Väänänen-Vainio-Mattila, K., Roto, V., Hassenzahl, M., 2008. Now let's do it in practice: user experience evaluation methods in product development. In: Extended Abstract of CHI'08. ACM, New York, pp. 3961–3964.

Vermeeren, A.P., 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. ACM, New York, pp. 521–530.