The trained panel method and its application in HCI research

ANONYMOUS AUTHOR(S)*

User interfaces utilising multiple modalities or even multisensory feedback are more common, creating the need for evaluation techniques that can take multiple quality dimensions under consideration. This paper demonstrates how the trained panel method can support the design and evaluation of physical or complex technological artefacts by mapping out design spaces based on their descriptive attributes. It is an expert-based method, and the goal is to derive a comprehensive description of a sample of existing artefacts or prototypes. The method entails training as well as multiple feedback sessions to ensure consensus among panel participants. We describe the advantages and limitations of the method by presenting how it was applied to identify salient attributes that are important in the design or evaluation of smartwatches. Apart from the specific case described in detail, we are also discussing how and in what context the trained panel method can provide value in HCI research and practice.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**.

Additional Key Words and Phrases: trained panel method, evaluation technique, expert technique, multisensory and multimodal artefact evaluation, AI evaluation technique

ACM Reference Format:

Anonymous Author(s). 2018. The trained panel method and its application in HCI research. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Our daily interactions with technology are not exclusively confined to the traditional visual user interface anymore. Instead, we are increasingly using devices with physical attributes, like mobile phones and wearable technology of various types. User acceptance of wearable technology is mainly defined by the functionality it provides, and the way we interact with it, as well as aesthetic considerations and quality attributes of its physical design [20]. In addition, devices are becoming increasingly multimodal by supporting various input and output modalities such as natural speech, touch, gaze, and gestures. At the same time, multisensory interfaces utilizing the chemical senses of smell and taste to interact with technology are actively explored [16, 26]. These developments highlight the need for techniques and methodologies that can be used to evaluate or inform the design of those multidimensional technological artefacts.

This paper aims to explore the value of the trained panel technique to evaluate or inform the design of physical, technological artefacts. We will demonstrate how a trained panel can identify critical design factors in specific domains through a case study. The trained panel technique is an adaptation of the Descriptive Analysis method, which is very popular and widely used in many scientific disciplines such as Sensory and Food science [8, 12], Marketing [25], and Audio Engineering [10]. However, in HCI literature, the technique can be encountered only rarely, typically in studies about website visual design [18, 19].

The proposed trained panel technique is an expert-based method whose utmost goal is to identify important design attributes. The panel participants help create and refine a comprehensive list of perceptual attributes based on some

Manuscript submitted to ACM

training stimuli. Multiple rating phases are implemented to provide feedback and calibrate panel performance since the method aims to reach a consensus about terminology and ratings. The process is analytical, and the attributes must be non-evaluative since only a description of important characteristics in which stimuli may vary systematically is the desired outcome. Hedonic or preference data from non-expert users can later be combined with the trained panel data using techniques such as Preference mapping [2].

Expert-based evaluation techniques such as expert reviews [9] or heuristic evaluations [15] are popular in HCI research. It is common practice to use experts in design workshops, focus groups, or the initial stages of questionnaire creation. During questionnaire development, it is very common to use a number of domain experts to brainstorm and generate a list of terms used later in exploratory factor analysis to identify latent dimensions. Even though all of those techniques are using experts, they vary considerably regarding purpose. Typically, in expert reviews, the goal is usability evaluation, and in design workshops, it is idea generation. In the method proposed in this paper, the goal is first to develop a common vocabulary among experts and afterwards to reach a consensus about the description of several artefacts.

Furthermore, the goal is not to generalize but to describe and provide a deep understanding of a specific domain. We consider this to be the most significant advantage of the technique. It is also why we believe it is highly appropriate for the characterization of complex multivariate artefacts. The paper's contribution is to demonstrate the advantages of this technique through a case study of smartwatches. We also provide a discussion about how this technique can fit into the HCI practitioners' toolkit.

2 METHOD

As aforementioned, the trained panel technique is based on a descriptive analysis method that originates in sensory and food science. Descriptive analysis is a methodology that provides quantitative descriptions of products based on the perceptions of a group of qualified panellists. During product description, all relevant sensations that are perceived are considered (e.g., visual, auditory, olfactory, kinesthetic) [23]. Typically, panellists are recruited for their ability to detect small differences in product characteristics and are part of the selection process of attributes used to assess the test stimuli. On some occasions, a set of predefined descriptors may also be given in addition to those derived in the panel discussions. Frequently, before the actual assessment, some additional products are utilised for scale calibration [11]. The purpose is to identify and quantify the intensities of the sensory characteristics of a product that are perceived (i.e., cannot be measured instrumentally) and are not affective (i.e., they do not entail the personal preferences of the assessors). Various multivariate statistical techniques are used to give feedback through the assessment phase [13] and summarise the results (e.g., attribute maps based on PCA). In the next section, we present details from a case study in which the trained panel technique has been used in the domain of smartwatches. Our aim is not only to provide an overview of the various phases of the trained panel method but also to showcase the outcome of applying this technique in the specific application domain

3 THE SMARTWATCH CASE

To illustrate the procedure and demonstrate the value of the trained panel technique in HCI research, we focused on the case smartwatches. The rationale for this choice was that this particular wearable technology has both a traditional visual user interface and physical characteristics. The case study presented here is part of a more extensive study to identify user preferences towards smartwatches [21]. The emphasis in this study was on user preferences, and trained panel data was only a small part of the analysis. Here we present the procedure we followed in-depth, providing details

Fig. 1. The six selected smartwatches in our sample: A) Polar m600, B) Motorola 360 2nd gen, C) Sony Smartwatch 3, D) Zeblaze Blitz, E) NO.1 G4, and F) NO.1 D6.

about the steps of the method and the decision made in each subsequent phase. In the following sections, we provide a detailed description of how we implemented the trained panel study to identify critical characteristics and how we mapped out a design space for the domain of smartwatches.

3.1 Participants

Our panellist recruitment goal was to select a group of people who would detect and conceptualise smartwatch design characteristics. Therefore, we tried to recruit people from a variety of disciplines who would have some type of previous experience with wearable technology. Our final panel study involved eight people; six were male and two females, aged 24-44 (M=32.75, SD=7.83). Their professional backgrounds were in visual design, interaction design, usability, techno-anthropology, manufacturing, and electrical engineering. We aimed for a group with diverse backgrounds to consider all possible aspects of the devices during the selection of attributes. However, since the goal of the trained panel study is to reach a consensus, an extensive training session has been held to ensure that our group had a shared understanding of each other's vocabulary.

3.2 Samples

Six smartwatches (labelled samples A to F in figure 1) corresponding to different brands were selected in the present study. Our selection strategy was based on including as much variety in design and quality characteristics as possible in a relatively small sample. Therefore, we selected smartwatches that varied on several factors such as prize, style, materials, colour, and shape. Since we also wanted to avoid potentially biasing factors such as brand and previous familiarity with the devices, we selected only android smartwatches with no clear brand insignia. For devices in which it was not possible, we carefully masked the brand logos with stickers. The descriptive profiles needed to be based only on attributes of the physical design or the user interface of the artefacts. The final set of smartwatches included in our sample can be seen in Figure 1.

3.3 Setting

The trained panel study was held in a round table format. The smartwatches were placed in the middle of the table, and each panellist could examine them as long as they needed. The panellists provided their ratings for the various smartwatch devices on individual laptops in front of them. The data collection was handled by a web application that was developed for this purpose. During session breaks, the study conductors used the specialised open-source application PanelCheck [24] to create visualisations of panel performance. Feedback about the performance was provided to the

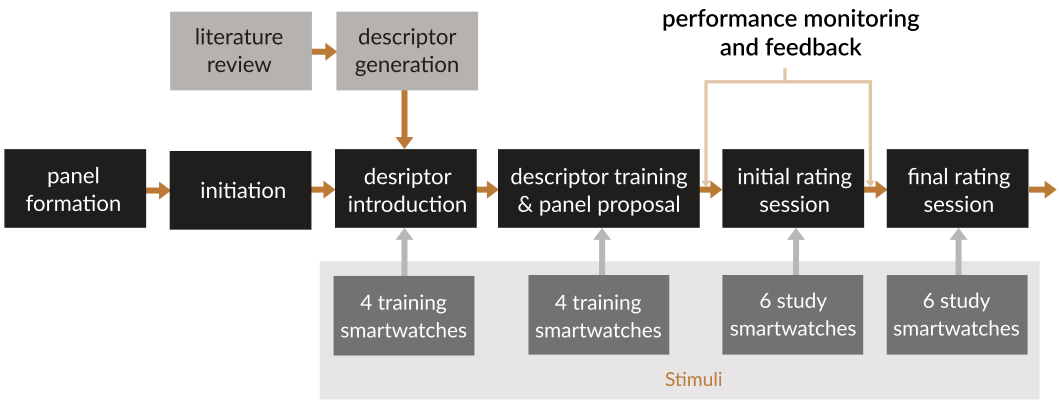

Fig. 2. Outline of the procedure followed in the trained panel study

panellists at the beginning of each new session by showing them several data visualisations on a big screen mounted in the meeting room wall. The selection rationale and purpose of these data visualisations will be explained in more detail in the following sections.

3.4 Procedure

The process followed in this study was comprised of five major stages. The outcome of the trained panel study was the creation of a comprehensive list of descriptors/attributes that could be used to profile smartwatches. This list was updated and modified throughout all the stages of the study. In the final stage, our participants used this list to rate the six sampled smartwatches on 100-point sliders.

At the beginning of the study, our panellists had to undergo a training procedure to refine the descriptor list and fine-tune panel understanding about it. The training session can be roughly subdivided into three major stages. In the first stage, participants were introduced to the goal and purpose of the study. In this phase, we emphasised that the panel's goal is to describe and not evaluate the smartwatches regarding personal likes and preferences.

In the second stage, participants were presented with an initial list of descriptors prepared beforehand by the authors through a literature review on smartwatches. These descriptors were then used to rate a training sample of four smartwatches (different than those in Figure 1). Panellists were encouraged to add a set of descriptors they considered as important for a smartwatch design or remove any of them. Subsequently, each of the identified descriptors was repeatedly discussed and defined. This process did not stop until all the panellists agreed about the importance of each descriptor and its definition. Afterwards, the descriptor list was tested in follow-up rating sessions.

Following the training, session participants were asked to rate the final sample of six smartwatches (Figure 1) twice. During the training and rating sessions, individual panellist's performance was monitored, and appropriate feedback was given were needed. The outline of the procedure can be seen in Figure 2. The whole process, including breaks and feedback sessions, took eight hours, and it was conducted in a single day.

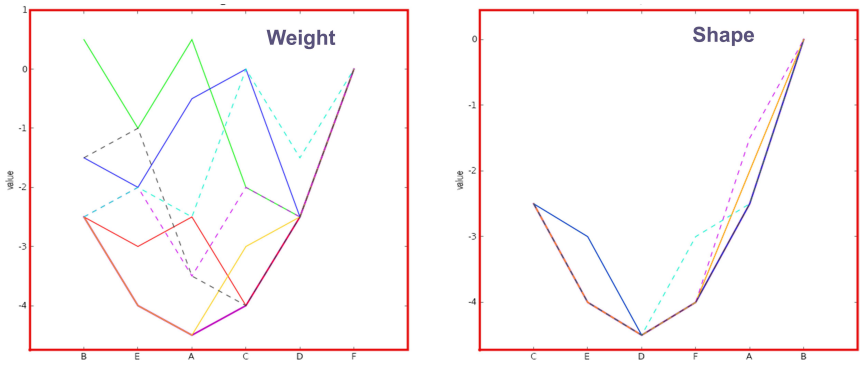

Fig. 3. Eggshell plots for the descriptors weight and squareness. For the descriptor weight, the considerable panel disagreement is clearly visible by the multiple "cracks".

4 ANALYSIS

Data analysis was conducted both during and after the panel session. During the session, the goal was to give feedback about the descriptors to the participants and help them reach a consensus. The ratings were used to provide feedback about descriptor and panel performance using appropriate data analysis and visualisation techniques. No single analysis method can provide sufficient inside into a panel performance. For this reason, several univariate and multivariate inspection techniques have been applied both for individual differences identification and for descriptor suitability assessment. Participant rating differences were explored through raw data visualisations such as histograms, box plots, principal component analysis, and eggshell plots [3].

To use eggshell plots [4] all individual assessor ratings for a descriptor have to be transformed into rankings and plotted alongside the consensus rank. The resulting plot's resemblance to an eggshell is the reason for its name [11]. Differences in the ranking order of the smartwatches among individual assessors concerning a certain attribute resemble cracks in the "eggshell". A large amount of cracks is a result of an increased disagreement between participants. In Figure 3, for example, the plot for the attribute weight has considerable more cracks than the attribute squareness. These differences are not a result of scale usage but of genuine disagreement among participants.

The next technique used was panel consonance plots [3] which are based on principal component analysis of each descriptor and for all participants. This is a multivariate method to identify panel agreement regarding individual descriptors. A well-calibrated panel should produce a unidimensional space which means that most of the variance can be attributed to the first component [8]. Having participants spread all over the chart indicates disagreement among panellists regarding how they use the attribute in question (see figure 3).

5 RESULTS

Two are the main results from our panel study: first, the list of descriptive descriptors that represent important characteristics in the design space of smartwatches. Second, the descriptive profiles that have been created by using this list to rate our sample devices. The outcome of the training session was a list of 36 descriptors and accompanied by a short description. In the final part of the session, each panellist rated the smartwatches in our sample on all identified descriptors. Examples of the identified attributes from the panel include noisiness, firmness, seriousness, bulkiness, etc.

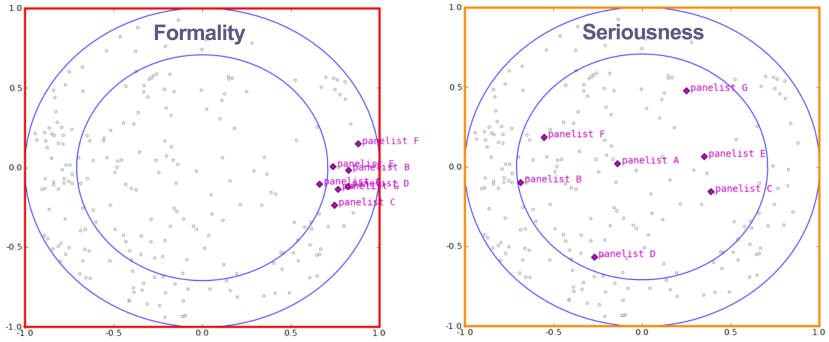

Fig. 4. Consonance plots for the descriptors formality and seriousness. The left plot shows general agreement among participants indicating that they all had a common understanding about this descriptor. On the right, the Seriousness plot shows general disagreement.

The next step was to assess each of the descriptors to eventually remove the ones that did not perform according to predefined criteria. The evaluation criteria were: descriptor discrimination ability and panellists' disagreement.

The reason that an identified descriptor may not discriminate among sample devices is most probably the fact that the six smartwatches did only differ slightly or not at all regarding that descriptor. The discrimination ability of the 36 descriptors was assessed through mixed-model ANOVAs with smartwatches as fixed factors and panellists as random ones. Since non-significant main effects for a descriptor usually means that there were no discernible differences in the sample, it is safe to remove it from further consideration. The next step was to examine whether some of the descriptors caused disagreement among the panellists. Significant participant and interaction effects could be an indication of panel disagreement regarding a specific descriptor. The mixed-model ANOVA analysis can give a first indication about which descriptors are problematic, but no single analysis method can provide satisfactory results on its own.

For this reason, several univariate and multivariate inspection techniques (e.g., Boxplots, Histograms, Profile plots, eggshell plots, Tucker-1 correlation plot) [11] have been applied to identify the descriptors that created considerable disagreement among panellists. At the end of this process, we removed three descriptors due to low discrimination ability since they had a non-significant main effect (p values ranging from 0.06 to 0.4). In addition, we removed another 11 descriptors since multiple methods confirmed that there was a lack of consensus in our panel about their meaning. The final list of 22 attributes is presented in Table 1.

Using participant ratings on these 22 attributes allowed us to create descriptive profiles for our six sample smartwatches. We created an attribute map to simplify results and visualise smartwatches' design space and their descriptive attributes (see Figure 5). This map is created by conducting Principal Component Analysis (PCA) on the data matrix consisting of stimuli in rows and attribute ratings in columns. The resulting map from this method is usually a low dimensional representation of the initial dataset. However, it can be used to reveal latent structure in panel perceptions and important attributes that significantly differentiate stimuli from each other.

Examination of this map reveals a design space that can be roughly divided into four quarters. Smartwatches on the top were perceived as more compact, more straightforward in design, and more similar to traditional watches, while the ones on the bottom looked more like smartwatches, were more complex, and were perceived to have more features. Devices on the left were perceived as more robust and expensive, and the ones on the left were flashier and noisier. Attribute proximity is an indication of correlation that can give a specific direction for design. For example, a

Descriptors	Description and scale
Shininess	How shiny the smartwatch is (glossy/matt)
Built Quality	The built quality of a smartwatch (fragile/robust)
Price	The perceived price of the smartwatch (cheap/expensive)
Style	The style of the smartwatch (sports/formal)
Size	The size of the smartwatch (bulky/compact)
Complexity	How complex in terms of design elements the smartwatch is (simple/complex)
Smartness	How much the smartwatch is perceived to be smart (watch/smartwatch)
Waterproofness	How much water-resistant the smartwatch is (non-waterproof/waterproof)
Attention	How much attention the smartwatch attracts (modest/flashy)
Watch-Noisiness	How noisy the smartwatch is when shaken (not-noisy/noisy)
Felt-Temperature	How does the smartwatch feel when someone wears it (cold/warm)
Prototypicality	How typical is the form of the smartwatch in relation to a wristwatch (non-typical/typical)
Shape	The shape of the smartwatch's face (round/square)
Colour	How colourful was the smartwatch's idle screen (colourless/colourful)
Brightness	How bright the smartwatch's idle screen is (dull/bright)
Resolution	How crisp the smartwatch's display is (grainy/crisp)
Swipe-Responsiveness	How responsive the smartwatch's display is when swiping (non-responsive/responsive)
Features	The number of features the smartwatch offers (few-features/many-features)
Bracelet-Traditionality	How traditional the bracelets' lock mechanism is (non-traditional/traditional)
Touch	How do the bracelets feel to the touch (harsh/soft)
Bendiness	How bendable the bracelet joints are (Rigid/flexible)
Button-Noisiness	How noisy the smartwatch's buttons are (when pressed) (not noisy/noisy)

Table 1. The final list of 22 descriptors able to describe the design space of smartwatches that resulted from the trained panel study

device was perceived to be expensive when it also had a crisp, high-resolution screen and was highly responsive to swipe gestures. On the other hand, flashy, noisy devices that were perceived to have many features were perceived as cheaper. It is also interesting to note that participants' perception of expensiveness did not necessarily correspond to the smartwatches' actual prizes.

6 DISCUSSION

In the previous sections, we described the process of applying the trained panel method in a specific case study. Then, we showed the steps we followed to train our panellists to gain a shared understanding and refine their vocabulary to describe the artefacts in the application domain of smartwatches. Finally, we presented the outcomes of this activity which are the list of domain-specific attributes accompanied by descriptions and scales and a profile of the six smartwatches based on these attributes. In this section, we will discuss how the trained panel method relates to other

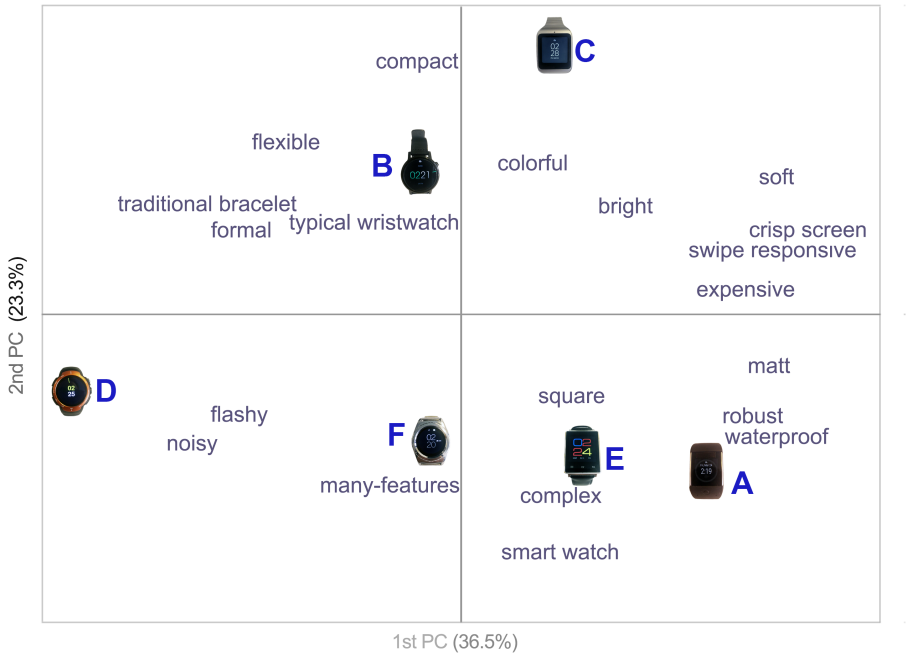

Fig. 5. Attribute-based map depicting the smartwatch design space.

expert-based methods. We will then outline the advantages and disadvantages of the technique. Lastly, we will present application domains and contexts in which we believe this technique can be valuable.

6.1 Comparison with other expert methods

As mentioned before, expert-based methods in HCI research vary considerably in purpose, outcomes and application areas. Experts are used in heuristic evaluation to assess compliance of a user interface with a set of predefined rules/heuristics. Those heuristics can be generic principles, commonly occurring errors, or guidelines for design for a specific quality dimension (e.g. usability [14]) or a particular application domain (e.g., speech interfaces [27]). The goal is to evaluate the design and improve the user interface on specific quality dimensions. In the panel method, the purpose is to get a non-evaluative description or a profile for a set of artefacts. Also, the descriptor list is not predefined and is developed every time to uniquely fit the application domain in question.

Another expert technique that can be encountered in HCI research is Delphi studies (e.g., [6, 22]). Delphi studies have many similarities with the trained panel method. It entails utilizing experts in an iterative approach of multiple phases of discussion and feedback until consensus is reached. However, the purpose of Delphi studies is not to describe and profile a set of artefacts, but it is to make future predictions and forecasts about the future of a specific technology or to extract guidelines from domain experts (e.g., [22]).

Another practice in which experts are involved in HCI research with striking similarities to the trained panel method is during the first stages of scale development. At the beginning of this process, it is very common to involve a diverse group of experts to generate items/adjectives that can describe the phenomenon or the quality the scale is trying to

assess (e.g., see [7]). Usually, a literature review is also held beforehand to establish an initial list of items given to the experts for review. We followed a very similar procedure in the initial stage of the trained panel method, as we have described in section 3.4. The main difference between those methods is that in scale development, experts only brainstorm and generate items. In contrast, in the trained panel method, experts are also the responsible to evaluate the fitness of the items in the subsequent stages. The purpose of a scale instrument is to capture the underlying structure in end-user perceptions regarding the quality of investigation. If, for example, the scale is developed to assess how users perceive the usability or the aesthetic design of a user interface, then the target of the investigation is potential users of the interface. Therefore a large number of potential end-users are involved in subsequent steps of exploratory and confirmatory factor analysis. In the trained panel data method, the purpose is not to gain insights into end-user perceptions. It is to achieve accurate profiles of the specific artefacts after identifying attributes in which they systematically differ. The list of descriptors identified cannot be transferred to another domain. The goal of scale development is to ensure that it can be applied to different examples within a domain. For example, the aesthetics scale [7] was developed to evaluate all types of websites. This is not the case for the trained panel method we described. It is doubtful that the list of descriptors we developed would be optimal to describe a new evolution of smartwatch design some years from now. The method's goal is to gain an accurate snapshot of a set of artefacts that can be used in later stages to be connected to end-user preferences (e.g., [19]), or to create a design space as in figure 5. Compared to using scales, using the trained pane method means sacrificing generalizability to gain accuracy.

6.2 Advantages and disadvantages

Before outlining the advantages and disadvantages of the trained panel method, we will summarize the main characteristics of the technique. First, it involves experts in the specific or relevant application domains. Second, participants are engaged in the creation and refinement of the descriptor list so that, at the end, only those that discriminate and are commonly understood are included. Third, the method follows an iterative process of discussions, ratings and feedback until consensus is achieved. Finally, after the training phase is finished, the participants create profiles for the artefacts in the sample by using a well-defined and commonly agreed-upon set of attributes.

The main advantage of this technique is that it provides certainty that participants understood and used the attributes in the same way. When we ask naïve participants in evaluation studies to rate something on a scale, they likely use different cues and strategies to accomplish the task. For example, if we ask them to rate how symmetrical an object is, they may subconsciously use horizontal, vertical, or even two-line symmetry cues to assess it. In the trained panel method, those differences can be identified during the training phases. Consonance plots, for example, can illustrate if there was disagreement about a specific descriptor and who deviated from group consensus. This information is used in feedback sessions in which it is decided if that descriptor should be removed or the list should be updated with better descriptions. For example, during the evaluation, we show high disagreement for the descriptor Noisiness. In the discussion, it became clear that some participants took the term metaphorically (i.e., Noisy design) while some others literally (how much noise it makes when it is rattled). Therefore, we decided to split the descriptor into two separate ones and update the list. During later stages, the Physical-Noisiness descriptor was split again into Button-Noisiness and Bracelet-Noisiness. However, two of three Noisiness descriptors did not discriminate significantly among our smartwatches and were removed from the final list. It should be noted that this approach would not be optimal if we were using affective or evaluative constructs instead of descriptive attributes. Variation in an evaluative contract could be capturing meaningful differences in participant perceptions. However, the method focuses on descriptive or collative attributes [1] that cannot be objectively measured but are not subjective either [17].

Another advantage of this method is its flexibility. Using established scales could be an alternative approach in some circumstances, but the results would reflect generic attributes that may or may not be important in a specific domain. To elaborate on this, we will use a movie analogy as an example. Let us assume we use a movie assessment scale to evaluate several movies. If this scale is generic enough to assess, for example, both horror and comedies, it will probably miss some factors that are unique to a specific movie genre (e.g. scariness). With the panel method, the factors emerge from the artefact sample, making the method adaptable to the specific domain.

However, the main disadvantage is that it is an intensive and time-consuming method to apply that can also become costly considering that it involves experts. As studies have shown [5, 28], if experts are not available, 'naïve' assessors can be used instead since they can also produce adequate results concerning the criteria of discrimination, consensus, reproducibility as long as they undergo more rigorous training. Another disadvantage is that it requires multiple artefacts, which can be a problem for new products if multiple prototypes are not available. Finally, for the method to produce good results, the artefacts have to be different to some extent. For example, it would be less optimal to perform a trained panel study with prototypes that differ only on a small number of design characteristics.

6.3 Application areas

The method can be applied to any application domain as long as several artefacts can be identified that are not identical and vary in some meaningful way. As mentioned before, the method can be encountered in HCI literature mainly in studies assessing visual or physical design (e.g., website design, wearable technology). However, we believe that it can be a valuable method for evaluating multimodal or multisensory interfaces such as ambient devices using the chemical senses of taste or smell. In addition, we believe that with some adjustments, the method could be used to evaluate concepts with no physical manifestation. For example, it could be used to profile the character or personality of voice assistants (e.g., [30]) based on audio assessment, voice tonality and content of responses. Or it could be used to assess the social media presence of institutions or companies (e.g. [29]) by selecting samples of posts from a specific period.

Besides those application areas, we also believe that the method can be valuable in designing and evaluating AI systems. It is common practice, for example, to use crowd-workers to annotated training data in supervised learning. The process has similarities to the trained panel method since the goal is accuracy and consistency. The trained panel method could be used to develop appropriate label descriptions or as a learning tool to train annotators. Finally, since the trained panel method can be applied using panellists with multidisciplinary backgrounds, it would be interesting to investigate if it could be adapted to accommodate AI system evaluation regarding fairness or bias.

7 CONCLUSIONS

In this paper, we demonstrated the advantages of the trained panel technique. We showed how we used panellists to identify important design characteristics of multidimensional devices such as smartwatches. During feedback sessions, we realised how easy it is for panellists even after extensive training to misunderstand the definition of an attribute. Fortunately, the technique entails multiple contingencies to ensure that only ratings that reflect a common understanding are used to create the final attribute space. This ensures that the final map is interpretable and reflects the panels' perceptions. These types of maps could be used to inform new design directions or refinement of prototypes in similar studies. In the future, we intend to combine hedonic data about user preference towards the devices to identify attributes that could be significant drivers of preference in the domain of smartwatches. Finally, we also intend to test the method with other types of technological artefacts to further assess the technique's value by testing it in other domains.

REFERENCES

[1] Daniel E Berlyne. 1973. Aesthetics and psychobiology. *Journal of Aesthetics and Art Criticism* 31, 4 (1973).

[2] Douglas J. Carroll. 1972. Individual differences and multidimensional scaling. *Multidimesional scaling. Theory and applications in the behavioral sciences, Theory* 1 (1972), 105–155.

[3] Tobias Dahl, Oliver Tomic, Jens P Wold, and Tormod Næs. 2008. Some new tools for visualising multi-way sensory data. *Food quality and preference* 19, 1 (2008), 103–113. https://doi.org/10.1016/j.foodqual.2007.07.001

[4] David Hirst and Tormod Næs. 1994. A graphical technique for assessing differences among a set of rankings. *Journal of Chemometrics* 8, 1 (1994), 81–93. https://doi.org/10.1002/cem.1180080108

[5] F Husson, S Le Dien, and J Pagès. 2001. Which value can be granted to sensory profiles given by consumers? Methodology and results. *Food quality and preference* 12, 5-7 (2001), 291–296.

[6] William Jones, Robert Capra, Anne Diekema, Jaime Teevan, Manuel Pérez-Quiñones, Jesse David Dinneen, and Bradley Hemminger. 2015. *"For Telling" the Present: Using the Delphi Method to Understand Personal Information Management Practices.* Association for Computing Machinery, New York, NY, USA, 3513–3522. https://doi.org/10.1145/2702123.2702523

[7] Talia Lavie and Noam Tractinsky. 2004. Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies* 60, 3 (2004), 269–298. https://doi.org/10.1016/j.ijhcs.2003.09.002

[8] Harry T Lawless, Hildegarde Heymann, et al. 2010. *Sensory evaluation of food: principles and practices.* Vol. 2. Springer.

[9] Robert L Mack and Jakob Nielsen. 1995. Usability inspection methods: Executive summary. In *Readings in Human–Computer Interaction.* Elsevier, 170–181.

[10] ville-veikko mattila. 2002. ideal point modelling of speech quality in mobile communications based on multidimensional scaling (mds). *journal of the audio engineering society* (april 2002).

[11] Tormod Næs, Per Bruun Brockhoff, and Oliver Tomic. 2011. *Statistics for sensory and consumer science.* John Wiley & Sons.

[12] Tormod Naes and Einar Risvik. 1997. Multivariate analysis of data in sensory science. *Chemometrics and Intelligent Laboratory Systems* 36, 1 (1997), 75–76. https://doi.org/10.1016/S0169-7439(96)00072-X

[13] Tormod Næs and Ragnhild Solheim. 1991. Detection and interpretation of variation within and between assessors in sensory profiling. *Journal of sensory studies* 6, 3 (1991), 159–177. https://doi.org/10.1111/j.1745-459X.1991.tb00512.x

[14] Jakob Nielsen. 1994. *Usability engineering.* Morgan Kaufmann.

[15] Jakob Nielsen. 1994. Usability inspection methods. In *Conference companion on Human factors in computing systems.* 413–414.

[16] Marianna Obrist, Alexandre N. Tuch, and Kasper Hornbaek. 2014. Opportunities for Odor: Experiences with Smell and Implications for Technology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14).* Association for Computing Machinery, New York, NY, USA, 2843–2852. https://doi.org/10.1145/2556288.2557008

[17] Eleftherios Papachristos and Nikolaos Avouris. 2009. The Subjective and Objective Nature of Website Aesthetic Impressions. In *Human-Computer Interaction – INTERACT 2009*, Tom Gross, Jan Gulliksen, Paula Kotzé, Lars Oestreicher, Philippe Palanque, Raquel Oliveira Prates, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 119–122.

[18] Eleftherios Papachristos and Nikolaos Avouris. 2011. The Application of Preference Mapping in Aesthetic Website Evaluation. In *Human-Computer Interaction – INTERACT 2011*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 616–619.

[19] Eleftherios Papachristos and Nikolaos Avouris. 2013. The Influence of Website Category on Aesthetic Preferences. In *Human-Computer Interaction – INTERACT 2013*, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 445–452.

[20] Anna Perry, Laura Malinin, Eulanda Sanders, Yan Li, and Katharine Leigh. 2017. Explore consumer needs and design purposes of smart clothing from designers' perspectives. *International Journal of Fashion Design, Technology and Education* 10, 3 (2017), 372–380. https://doi.org/10.1080/17543266.2016.1278465

[21] Dimitrios Raptis, Eleftherios Papachristos, Anders Bruun, and Jesper Kjeldskov. 2020. Why did you pick that? A study on smartwatch design qualities and people's preferences. *Behaviour & Information Technology* 0, 0 (2020), 1–18. https://doi.org/10.1080/0144929X.2020.1836259

[22] Kai Riemer, Judy Kay, et al. 2019. Mapping beyond the uncanny valley: A Delphi study on aiding adoption of realistic digital faces. In *Proceedings of the 52nd Hawaii International Conference on System Sciences.* https://doi.org/10.24251/HICSS.2019.577

[23] Herbert Stone and Joel L. Sidel. 2004. Special Problems. In *Sensory Evaluation Practices (Third Edition)* (third edition ed.), Herbert Stone and Joel L. Sidel (Eds.). Academic Press, San Diego, 279–335. https://doi.org/10.1016/B978-012672690-9/50012-3

[24] Oliver Tomic, Giorgio Luciano, Asgeir Nilsen, Grethe Hyldig, Kirsten Lorensen, and Tormod Næs. 2010. Analysing sensory panel performance in a proficiency test using the PanelCheck software. *European Food Research and Technology* 230, 3 (2010), 497–511. https://doi.org/10.1007/s00217-009-1185-y

[25] Ellen Van Kleef, Hans CM Van Trijp, and Pieternel Luning. 2006. Internal versus external preference analysis: An exploratory study on end-user evaluation. 17, 5 (2006), 387–399. https://doi.org/10.1016/j.foodqual.2005.05.001

[26] Carlos Velasco, Marianna Obrist, Olivia Petit, and Charles Spence. 2018. Multisensory Technology for Flavor Augmentation: A Mini Review. *Frontiers in Psychology* 9 (2018), 26. https://doi.org/10.3389/fpsyg.2018.00026

[27] Susan Weinschenk and Dean T Barker. 2000. *Designing effective speech interfaces.* John Wiley & Sons, Inc.
[28] Thierry Worch, Sébastien Lê, and Pieter Punter. 2010. How reliable are the consumers? Comparison of sensory profiles from consumers and experts. *Food quality and preference* 21, 3 (2010), 309–318.