

Measuring Desirability: New methods for evaluating desirability in a usability lab setting

Joey Benedek and Trish Miner

Microsoft Corporation, 1 Microsoft Way, Redmond, WA 98052

joeyb@microsoft.com and trishmi@microsoft.com

Abstract

Difficulty can arise when a practitioner wants to get user input on intangibles such as “desire” and “fun” in a usability lab setting. This paper will introduce you to methods we’ve created to collect feedback on “desirability” and give some background on how we developed them.

Introduction

Formal usability tests in a lab setting are an excellent tool to evaluate whether users can complete tasks; however, the technique has not been as effective for measuring intangible aspects of the user experience such as “fun,” “enjoyment,” or whether the product is desirable enough to purchase. One standard method used to evaluate these intangible aspects is a questionnaire with Likert scales. One problem with this method is that the topics of the questions or anchors on the scales are assigned by the practitioner and often do not mean as much to a participant brought into the lab. In addition, there can be a tendency to give very similar and often positive ratings for each question in the lab environment. Another option is an interview. While this can result in useful data, this approach can be time consuming and with some users it can be difficult to elicit their candid or more negative feedback. In addition the subjective data that is gathered can be challenging to analyze. Because of these barriers, among others, usability engineers in our organization are unlikely to assess “desirability” while evaluating products in the usability lab.

Since product team members had expressed an interest in knowing more than “is the product usable,” we decided to explore ways to measure intangible aspects of the user experience quickly and easily in the lab. Initially this involved interviewing people across the company who have conducted research related to desirability of Microsoft products and brainstorming with them. This paper includes information about the brainstorm session used to generate new ideas, specifics about two different methods we developed as a result of the brainstorm session, and our process using and refining those methods.

The Brainstorm Session

To jumpstart our thinking we hosted a brainstorming session with 8 usability practitioners, including individuals who primarily conduct lab studies and individuals who conduct field research. The objective was to think beyond the traditional techniques we use to collect feedback, restrictions, limitations, and to elicit innovative ideas. After being introduced to the ground rules for the brainstorm session (e.g. withhold judgment, quantity of ideas not quality, encourage wild ideas) the group did a practice brainstorming exercise on “ideas for new television shows” to stimulate thinking. After the exercise, the participants received a written copy of the statement below to guide the brainstorming.

“You are a usability engineer and you want to find out what makes something desirable. What can you do?”

Desirable can be defined as, ‘Worth having or seeking, as by being useful, advantageous, or pleasing’ or you can replace desirable with words, such as those that follow, to stimulate ideas - fun, engaging, natural,

enjoyable, valuable, essential, suitable, worthwhile, beneficial, likeable, in demand, amusing, and appealing.”

The participants spent the remainder of the meeting brainstorming on the main topic, while various participants noted the ideas on large sheets of paper posted throughout the room. During the session the facilitator occasionally suggested different questions to think about to stimulate new ideas. Also, toward the end of the session each participant in the room voted on the 3-5 ideas they felt were best to pursue further.

After the brainstorm session we met to transcribe the brainstorm ideas and to select a few ideas to develop further. The criteria used to select the ideas were that the method be “quick” to administer and that the resulting data be easy to analyze. We did not base our decision on whether we could obtain “statistically significant” results. The two ideas we focused on developing, that are further described below, had the advantage of allowing users to choose words they would use to describe a product to a practitioner. It is important to note that all of the brainstorm ideas had some basis in prior knowledge of the participants and could be linked back to prior research in a variety of fields. To our knowledge the specific tools we developed have not been created before; however, the general concepts behind them are not new. Due to a variety of resource and time constraints we did not conduct a literature search on these topics.

The Faces Questionnaire

One idea was to use ambiguous stimuli, in this case different photographs of faces, as the basis of a questionnaire. The idea was inspired by a large body of work on facial expression and emotion, including the work of Paul Ekman[1]. The goal was to elicit feedback about the product, particularly emotions that arose for the participants while using the product (e.g. frustration, happiness). Our first draft had pictures of 6 different faces representing a wide array of emotions from joy to anger with a scale from 1-7 where 1 represented “Not at all like this” and 7 represented “Very much like this.” The participants were asked to circle the rating on the scale that most closely reflected how performing the tasks with the product made them feel. We expected the users to look at each face and decide what the face represented. They would then state what the face meant to them and rate the product based on that dimension using the Likert scale. While describing the face and rating, the usability practitioner probed to obtain more details about why they picked that particular rating, obtaining additional information about their reaction to the product. We anticipated that the participants might choose different words to describe the same face, so that it would be possible for each participant to be rating a slightly different emotion. This was a desired effect because we did not want average ratings, we really wanted to see what concepts emerged as descriptions of the product and to understand more from the participants about the characteristics of the product they felt contributed to their reactions. Unfortunately, the first iteration of the scale had facial expressions that were somewhat extreme (e.g. furious, ecstatic) and many participants felt they did not apply and this limited the feedback generated in response to the questionnaire. In the second iteration we used 6 photographs with more ambiguous facial expressions (See Figure 1 for an example from the revised questionnaire). These positive, negative and neutral faces did elicit more varied and less extreme descriptions from the participants, such as “frustration” rather than “anger.” While the questionnaire worked very well with some people, **users with less developed vocabularies found it difficult to describe the facial expression** and thus had difficulty picking a rating and discussing how their rating applied to the product.

Figure 1: Example item from the revised Faces Questionnaire

Look at the picture and circle the rating on the scale that most closely reflects how performing the tasks today made you feel.



Not at all
like this

1 2 3 4 5 6 7

Very much
like this

Product Reaction Cards

The other tool we developed is a large set of word cards that form the basis for a sorting exercise and more importantly discussion about the product. While there is a large body of literature on card sorting, recognition, and memory, to our knowledge the specific cards and method we developed for the usability lab are unique. The main advantages in our mind were that this technique does not rely on a questionnaire or rating scales and users do not have to generate words themselves.

After obtaining a large set of potential words from prior research, marketing materials, and team brainstorming, a set of approximately 75 words was compiled. Since there is a bias to give positive feedback in the usability lab, we made sure that at least 40% of the set consisted of negative words and phrases and tried to make the set cover a wide variety of dimensions. Each word was placed on a separate cue card and the set was given to the participant at the end of usability evaluations. The participant was asked to pick the words that best describe the product or how using the product made them feel. Once the user picked the words the practitioner returned to the room, recorded the words selected, and asked the user to narrow their set down to the 5 best. The practitioner then asked for details about why they picked each of the top 5 words. During this exchange the participants reveal a great deal of information including specifics about their interaction with the product as well as their reaction to the product concept and design.

This method was piloted in four separate lab studies, with different practitioners, and was very successful and popular for many reasons. From the usability practitioner perspective, a great deal of useful data is obtained in a short period of time, including details about what users liked and disliked, as well as details about their interaction with the product not previously shared during the session. The team enjoys coming to the lab to hear the information and leaves motivated to further improve the product. The participants in studies also spontaneously say that they enjoyed the discussion and sharing their feedback.

During the pilot testing all of the teams used the initial set of cards, and occasionally added their own cards to tap into different concepts. At the end of piloting there were a total of 64 additional candidate words. To make sure the base set covers key concepts that users think are important while not including unnecessary cards, we collected data from 16 end users who had never seen the product reaction cards. Each participant sorted the product reaction cards in two ways during the session. The first was a classic free sort mainly to identify words that may have been too similar. We then explained the purpose of the product reaction cards to each participant and asked them to sort the cards into 4 different groups based on how important it was for each word to be in the set. We used this feedback, as well as data on cards that had not been chosen in prior evaluations with the toolkit, to reduce the set. Originally we targeted a 60% positive and 40% negative/neutral balance. Since we had data to suggest that users choose

negative words during usability evaluations at that balance point, we decided to get as close to that level as possible for the final toolkit. Our data also suggests that with a toolkit of 120 cards the exercise can be completed fairly quickly. While it may have been possible to increase the size of the kit, we decided to stick as close as possible to 120 cards. Therefore we balanced the set by adding some additional negative words to have the same ratio we started with and the result was a final set that consists of 118 cards (See Figure 2 for the complete list of words). This final set has been used in 3 studies so far and is working very well.

Figure 2: Selected words from the set of 118 product reaction cards

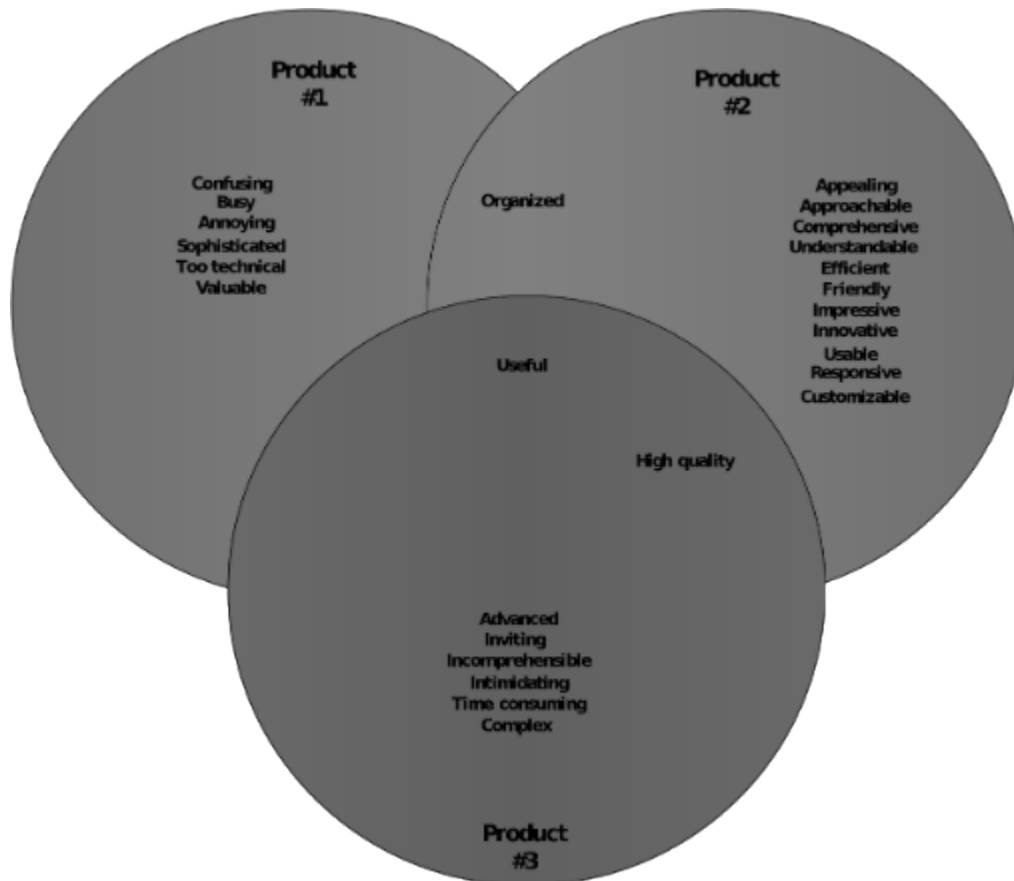
Accessible	Desirable	Gets in the way	Patronizing	Stressful
Appealing	Easy to use	Hard to use	Personal	Time-consuming
Attractive	Efficient	High quality	Predictable	Time-saving
Busy	Empowering	Inconsistent	Relevant	Too technical
Collaborative	Exciting	Intimidating	Reliable	Trustworthy
Complex	Familiar	Inviting	Rigid	Uncontrollable
Comprehensive	Fast	Motivating	Simplistic	Unconventional
Confusing	Flexible	Not valuable	Slow	Unpredictable
Connected	Fresh	Organized	Sophisticated	Usable
Consistent	Frustrating	Overbearing	Stimulating	Useful
Customizable	Fun	Overwhelming	Straight Forward	Valuable

The information gathered using the product reaction cards can be presented in a variety of ways, depending on the objectives of the evaluation. Figure 3 illustrates how the information from an early iterative study could be presented and Figure 4 shows how the data collected during a summative evaluation of different products or different versions of the same product could be compared. Another interesting alternative is to have team members use the product reaction cards to describe the product they are designing and to compare that with the words that users of the product choose as important. These results can help the product team describe their vision in a concrete way and to see how that relates to what the users of the product are interested in and may result in design changes to align those visions more closely.

Figure 3: Abbreviated example data from an early iterative study

Top 5 Cards Chosen	User Segment	Comments
Attractive	Beginner	Legible, clear, nice to look at, items are not fighting for attention, no flashing
	Intermediate	Color selection and icons
Efficient	Intermediate	After customizing, places can be reached with ease
Busy	Advanced	Page has too much stuff on it.
Frustrating	Intermediate	Changing Password. Didn't understand the words that were used.

Figure 4: Example data from a study comparing 3 different products. The same presentation style could also be used to compare different versions of the same product.



The Desirability Toolkit

In this paper we present 2 methods that currently form the basis for our Desirability Toolkit that is in use by usability engineers at our company. It is important to point out that these tools have a variety of limitations and this work is only a start toward measuring “desirability” and other intangibles in a usability lab setting.

Some limitations we want to acknowledge include the following:

- *We have intentionally created a practitioners tool.* These techniques are quick and limited. For example, the Faces Questionnaire is not designed to be a valid, consistent, measure, and we have intentionally traded that rigor for richer responses from the participants in the usability lab.
- *These techniques result in qualitative rather than quantitative data.* We do not look for statistical significance in the results of these tools. We do not provide ratings based on the Faces questionnaire and any counts of the cards in the product reaction set are only to indicate trends. The most important data from both tools is from the discussion with the participant in the study to determine their reaction to an item (e.g. whether they think it is positive or negative) and how they apply that item to the product being evaluated.

- *The results from these techniques cannot be generalized.* These techniques are not designed for making broad statements about all potential users of the product; they are biased toward information that we can use to judge the quality of the user experience for the participants who are in our usability evaluations and to suggest design changes.

There are a number of benefits of the tools we have developed and of measuring intangibles such as “fun” or “desire” in the usability lab:

- *Usability participants enjoy the toolkit exercises.* The participants in our labs like the desirability toolkit exercises, giving us much more positive feedback about the experience than when we use questionnaires and other similar techniques.
- *Candid feedback is obtained.* With prior techniques it has been difficult to break through social barriers to encourage the participants to share what they disliked or might change about the products they evaluate. In contrast with the toolkit the participants volunteer lots of information they may not have revealed previously. They become very engaged in the discussion and are willing to share negative feedback.
- *Quick to administer and to analyze data.* Usability practitioners collect useful data with these tools in as little as 5-10 minutes, recording data right into a form that can be presented to a product team.
- *The product team internalizes the users’ message.* Team members who have watched the toolkit exercises leave with a broader picture of how users are responding to the product. In addition to knowing whether users can accomplish tasks and where they run into problems, the team also knows what aspects of the design they react positively toward and what emotions people using their product experience and why. The team also seems to identify more with what the users have to say and leave motivated to improve the product.

While we will be continuing our work by refining these techniques and developing additional measures for the toolkit, we also want to encourage other practitioners to incorporate techniques to measure desirability in their evaluations and to share their results.

Reprinted with permission, Usability Professionals' Association, 2002
Conference Proceedings.

[1] A summary of Paul Ekman’s work can be found on his website at <http://www.paulekman.com>. One recent publication that may be of interest is Ekman, P. (1999) *Facial Expressions*. In T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion. Pp. 301-320. Sussex, U.K.: John Wiley & Sons, Ltd.