A METHOD FOR MEASURING USER PREFERENCES IN INFORMATION SYSTEMS DESIGN CHOICES

Complete Research

Martin Matzner, University of Muenster, ERCIS, Münster, Germany,
martin.matzner@ercis.uni-muenster.de

Moritz von Hoffen, University of Muenster, ERCIS, Münster, Germany,
moritz.von.hoffen@ercis.uni-muenster.de

Tobias Heide, University of Muenster, ERCIS, Münster, Germany,
tobias.heide@ercis.uni-muenster.de

Florian Plenter, University of Muenster, ERCIS, Münster, Germany,
florian.plenter@ercis.uni-muenster.de

Friedrich Chasin, University of Muenster, ERCIS, Münster, Germany,
friedrich.chasin@ercis.uni-muenster.de

Abstract

Information System Design (ISD) applies information technology to achieve desired ends in organizations and implies many technology choices to be made. A successful design of information systems addresses the different views of all its stakeholders in these decisions. If we consider that sub-part of an IS that is intended to assist in customer processes, a purposeful assessment of the preferences of this anonymous mass is needed. Methods of Human-Centered ISD are not sufficient in that case for that they require too close integration of the subjects; and state of the art preference measurement techniques are likely to be too time-consuming and cognitively challenging if the number of alternatives is large. Building on the Q-Methodology, originally developed to reveal subjectivity in psychology, we suggest a novel method for user preference measurement. We report on a case in which we failed by applying standard techniques for user measurement, but succeeded with Q-Sort. By means of an experiment we subsequently compare the mentioned methods and identify root causes for failure and success we experienced in the case, which for Q-Sort include short execution time, measuring many design choices at one time, satisfaction of the interviewees, and an effective IT support.

Keywords: Design Choices, Information Systems Design, Psychometric, Preference Determination.

1 Introduction

Information Systems Design (ISD) is concerned with the application of information technology (IT) in organizational settings (Hirschheim et al., 1991). ISD can be described as the search process used to find an effective (IT-based) solution to an organizational problem using available means (the technologies)—including making decisions on either using or not using certain technologies to achieve desired ends (Hevner et al., 2004; Simon, 1996). Drawing on the social construction of technology (Bijker, 1987; Howcroft et al., 2004), scholars emphasized that problem space and solution space emerge in co-evolution (Dorst and Cross, 2001). Both spaces converge while the IS' stakeholders (such as managers, analysts, and users) repeat sequences of design, sense-making, and negotiation (Becker et al., 2013; Lyytinen et al., 2008). This is why stakeholders rely on meaningful responses from each other (Hirschheim et al., 1991).

In the following we focus on potential customers of an organization as specific user group, and we consider ISD activities related to the front stage of the IS which summarizes means intended to be used by these customers. To this end, Human-Centered ISD proposed participatory design and interaction design to represent user-worlds in the design process (Gasson, 2003). “Participatory design” summarizes methods that facilitate negotiation, shared construction, and collective discovery (Müller and Kals, 2004) in ISD. “Participatory design” examines the ways potential users work with technologies from the solution space. Both approaches alike rely only on the voices of a very few “representatives” (Gasson, 2003) and require for close integration and personal involvement into the ISD process (IDEO, 2011) which is hard to achieve in case of the anonymous mass of an enterprise’s customers.

Questionnaire-based surveys (Abras et al., 2004) can mitigate the problem of inadequate user representation. However, assessing preferences in this way is both methodologically and operationally challenging. Several possible means from the solution space have to be assessed at one time. Questionnaires constructed with state of the art preference measurement techniques are thus likely to be lengthy and cognitively challenging and are also likely to exhaust the interviewees. Against this backdrop, in this article we report how we failed in creating a questionnaire for preference measurement while we developed an innovative IS for electric vehicle charging. Our failure using state of the art techniques motivates the following research question: *“How can user preferences with regard to design options from the solution space be effectively measured in ISD?”*

The contribution of this paper lies in the development of a method that worked better for us. This method builds on Stephenson’s (1953) *Q-Sort*. We executed an experiment which demonstrates that the methods stands out in situations in which: a large range (7+) of options needs to be assessed; each option is assessed by a variety of factors; the participants’ preference orders vary strongly in the middle section; time is critical and the assessment has to be taken forward efficiently and promptly; participants’ satisfaction is relevant.

The remainder of the work is structured as follows: Section 2 introduces the research background on design choices in ISD and related work on user preference assessments in order to situate our method within the existing body of research. Section 3 introduces our project setting, and section 4 introduces our attempts to measure user preferences. Next, we describe the experiment we conducted to compare methods for measuring preferences. Finally, we conclude the paper with an outlook to the future and address identified limitations.

2 Research Background

2.1 Design Choices in ISD

When architects design the technology set of an IS, certain design choices must be made at various points in the process. Some of these design choices affect the user, as they directly influence interactions with the user or are core decisions that determine the artifact’s functionality.

The popular design science research methodology (DSRM) requires that one makes the right design choices in order to build good artifacts (Hevner et al., 2004). These decisions are located in the DSRM process in the design development step (Peffers et al., 2006). Design choices are design alternatives that follow similar goals but are realized differently and may be conflicting. They implement design objectives, which should be maximized (Pottie, 1995).

A developer can either trust his or her instincts or base the decision on a theoretical foundation. A developer may also consider similar, successful products and adopt their best-practice design decisions or ask target users which choice they prefer.

All of these possibilities have certain advantages and disadvantages. Trusting one’s instincts is a way to deal with design choices quickly, but it is extremely subjective, and choices made in this way have no rational foundation. A more objective decision on design choices relies on accepted theories and building an argumentational foundation (Clegg, 2000). Steps for building strong arguments include deriving choices, managing the sensitivities of choices and solutions, evaluating the options' influence on the design, and establishing a design rationale (Bate and Audsley, 2004). Two alternatives are the Architecture Trade-Off Analysis Method (ATAM) (Clements et al., 2001) and a method that derives assessment criteria using Goal Question Metrics (GQM) (Basili and Rombach, 1988). Most design-choice problems in information systems design and development use such foundations to reduce the risk of making the wrong decision. However, users have often not behaved as expected (Malhotra and Galletta, 2004). By adopting best-practice choices from already successful products, designers can be more confident; the closer the new artifact is to the adopted artifact, the greater the chance that the design choice will also be successful. In other words, similarity minimizes the risk. However, from a competitive point of view, it is not a good idea to adopt all design choices from a similar product, as doing so reduces the design's innovativeness and unique selling points. On the other hand, asking target users will lead to answers regarding design choices that are exactly appropriate for the artifact at hand without losing competitive advantage or cloning existing approaches. However, asking users is the most complex approach, and there is a risk of choosing the wrong set of users or asking the wrong questions. This work focuses on the approach of asking the users.

2.2 Methods for Assessing User Preferences

There are numerous ways to assess user preferences partially originating from the scholarly discourse on requirements engineering, in which several conceptual approaches and techniques have been brought forth (Pohl, 2010). A popular technique for measuring preferences is *Conjoint Analysis*, which estimates preference structures through evaluation of a set of alternatives with a specified combination of attributes (Green and Srinivasan, 1990; Srinivasan and Park, 1997). The problem with Conjoint Analysis is that it does not work well for a large number of attributes—seven or more—as it burdens the respondent with information overload (Green and Srinivasan, 1990). The *Self-explicated approach*, an alternative method for preference structure measurement, enjoys a similar level of popularity among researchers (Sattler and Hensel-Börner, 2003). The Self-explicated approach questions the respondent separately on each attribute, thereby minimizing the information-overload problem (Srinivasan and Park, 1997). In addition to the superior handling of a large number of attributes, the Self-explicated approach offers, among other things, greater usability for respondents, especially concerning their cognitive effort, and greater ease in terms of data collection and analysis (Sattler and Hensel-Börner, 2003).[1] From the various methods of the Self-explicated approach, we chose *Ranking, Rating, Maximum Difference Scaling, and Q-Sort* to be tested in our experiment. The choice of *Maximum Difference Scaling* and *Q-Sort* is based on the evolution of methods used in the first and second research cycle, as is described in section 4. We added *Ranking* and *Rating*, as they are standard methods for preference measurement, and they provide a good benchmark for the usability of the method because of their simplicity.

2.2.1 Ranking

In a simple preference ranking, the respondent is asked to rank the attributes from most attractive to least attractive according to his or her personal preference. The advantage of *Ranking* is that each scale point is used only once, and the resulting data is ranked ordinally (Cohen, 2003). With an increasing number of attributes, the *Ranking* method's usability for the respondent decreases, as the procedure becomes cognitively challenging and time-consuming (Alwin and Krosnick, 1985; Chrzan and Golovashkina, 2006; Munson and McIntyre, 1979; Rokeach, 1973). Other disadvantages of *Ranking* are potential order effects and the lack of ties and absolute scores (Cohen, 2003). A use case of *Ranking* for preference measurement

[1] For an extensive discourse on the advantages and disadvantages of conjoint analysis versus the self-explicated approach and a comprehensive overview of studies that compare the self-explicated approach and conjoint measurement, see Sattler and Hensel-Börner (2003).

in IS is described by Prieto-Diaz and Freeman (1987), who used the method to find functionally equivalent components for software reuse.

2.2.2 Rating

Using the preference-rating measurement, the respondent rates each item on a scale, e.g. from 1 (least attractive) to 100 (most attractive). This simple task has the disadvantage of allowing respondents to rate every choice equally high, in which case the rating does not fully reflect their choice. This process may also introduce a social desirability bias, when respondents are free to rank highly even though items they do not value but they believe others do, or an awareness bias, when respondents rank highly only those items they understand (Bacon, 2003). However, the *Rating* method is highly usable for the respondents because of its simplicity, and it has none of the major disadvantages of the *Ranking* method (Alwin and Krosnick, 1985). Even so, its simplicity might reduce the quality of the data, as *Rating* requires less effort from the respondent because it does not force him or her to differentiate among attributes (Alwin and Krosnick, 1985; Feather, 1973).

2.2.3 Maximum Difference Scaling

Maximum Difference Scaling (MaxDiff), introduced and developed by Louviere (Finn and Louviere, 1992; Louviere, 1991; Louviere et al., 1994), extends the method of paired comparison, where the preferred attribute is chosen from a pair of attributes (Thurstone, 1927). The respondent chooses a combination of the most and least preferred attribute from choice sets that contain only four to six of the total number of attributes available in the sample (Chrzan and Golovashkina, 2006; Cohen, 2003). In addition to a traditional Conjoint Analysis, *MaxDiff* forces not only intra-item comparison but also inter-item comparison of levels (Cohen, 2003). Comparing *MaxDiff* to other preference-measuring methods such as *Q-Sort* and *Rating*, Chrzan and Golovashkina (2006) found a high level of performance in inter-item differentiation and predictive validity for *MaxDiff* at the price of a long task length for the respondent. Cohen (2003) matched *MaxDiff* against the *Rating* and paired comparison methods and advocated *MaxDiff* because of its ease of use for both respondents and researchers and because it is scale-free and, therefore, easy to compare. Lansing et al. (2013) used the method in an IS context to measure consumers' preferences in cloud services.

2.2.4 Q-Sort

The *Q-Methodology* (Stephenson, 1953), originally developed as a way to reveal subjectivity in psychology, has been widely adopted in the social sciences (Brown, 1996). In the Q-Methodology individual rankings are subjected to factor analysis in order to reveal correlations between personal profiles, which then indicate similar viewpoints or subjectivity (Brown, 1993). Using the *Q-Sort* technique, the respondent ranks the items in a given pattern of a quasi-normal distribution from most to least important (Chrzan and Golovashkina, 2006). This pattern consists of classes that represent different levels of importance, so it forces the respondent to rate items within the same class with the same level of importance but to distinguish between classes (Eckert and Schaaf, 2009). An unforced *Q-Sort* which does not enforce a normal distribution is possible as well but is not as widely used since the forcing characteristic is considered an integral part of the method (Müller and Kals, 2004). One advantage of *Q-Sort* is the possibility to have a relatively large number of attributes. Stephenson (1935, 1936a,b) used up to sixty attributes to measure preferences with *Q-Sort*. Segars and Grover (1998) provided a prominent application of *Q-Sort*, using the method to define a theoretically derived construct space for strategic information systems planning success. Thomas and Watson (2002) presented *Q-Sort* for management information systems and provided extensive examples of design, administration, and data analysis as well as an online application for the method.

3 Project Setting

The need to evaluate certain user preferences originated in *CrowdStrom* (Matzner et al., 2015), a research project on e-mobility, where we developed a business model for sharing and using privately owned e-mobility charging stations. The project included a mobile app and a web portal. During the development, several design choices had to be made, including the form of user authentication at the charging point and the payment methods to be offered. Both are part of the user's interaction with the developed artifact, so these choices directly influence the user's acceptance of the final product. In determining the most appropriate authentication and payment method, we decided to ask potential users which design choices they preferred in this specific scenario.

Before asking potential users for their preferences, one must identify relevant design choices and the target user group. In our case we identified the design choices to be tested by evaluating similar solutions and theoretical work from the fields of authentication and payment. Examples for competing design choices in this scenario are authentication with the personal mobile phone vs. authentication by chip-card and payment by credit card vs. monthly invoice. In all, we compared eight methods for authentication and eight methods of payment. Target users for using and sharing e-mobility charging points were car owners from the targeted region.

4 Assessing User Preferences in an ISD Project

4.1 First Research Cycle: Failure

4.1.1 Action-planning

During the first research cycle, we considered three alternatives for measuring preferences: *Ranking*, *Rating*, and *MaxDiff*. We weighed the strengths and weaknesses of each of these techniques, noting the advantage of the *Ranking* technique in providing data that is easy to work with but also the fact that, with the *Ranking* approach, we could not tell if respondents were indifferent to any of the alternatives. *Ratings*, on the other hand, would allow respondents to represent their indifference toward one or more alternatives, but respondents have a tendency to indicate all options are equally or nearly equally preferred (Louviere, 1991; Louviere et al., 1994). Therefore, we initially decided to use the *MaxDiff* method because of the ability to forcing respondents to make choices between alternatives while still representing the relative importance of the alternatives being rated.

4.1.2 Action-taking

We operationalized the method in the form of a questionnaire that consisted of an introduction to the project scenario, instructions on how to use *MaxDiff*, and two scenarios in which to apply the method in the domain of electric-vehicle-charging infrastructure: authentication methods and payment methods.

In each scenario, a total of eight authentication or payment alternatives were integrated into eleven choice sets, each consisting of four alternatives. Table 1 illustrates an exemplary *choice set*. Therefore, we planned for the respondents to assess a total of twenty-two choice sets in order to deliver his or her preferences in regard to the authentication and payment alternatives when charging an electric vehicle at a charging station. To ensure the applicability of the technique in our project setting, we performed a pre-test with eighteen junior and senior scholars from the IS field and eight IS students. The age of respondents ranged from 21 to 37 years, with 24 of the 26 respondents being male. The pre-test results questioned the applicability of the method in practice.

Best	Method	Worst
	Chip Card	
✓	Fingerprint Scanner	
	Chip Card + PIN	
	Log-in + Password	✓

Table 1. Exemplary choice set for authentication method.

4.1.3 Evaluating

The pre-test revealed five major potential barriers to respondents when they use the *MaxDiff* method:

1. The time required to read the instructions and fill out the choice sets ranged from ten to twenty minutes and was generally perceived as "way too long," resulting in half of the questionnaires' being filled out incompletely.
2. Respondents felt "challenged" by the method, as it forced them to identify the best and the worst alternative also in those sets where they felt indifferent towards all alternatives.
3. Several respondents reported using simplified techniques to speed up the process, such as picking their personal most and least favorite alternatives, marking them in the corresponding choice sets first, and then filling in the rest more or less randomly.
4. Most of the respondents reported being unsure about whether they were consistent in their choices.
5. Nearly every respondent reported being "frustrated" by the method.

4.1.4 Learning

The pre-test results showed us that, in order to perform a large-scale test, we needed to find a method that has the major advantages of the *MaxDiff* method while also being less time-consuming, allowing for indifference toward alternatives, providing results that are comparable with the results that can be achieved with other techniques, and most importantly, result in both satisfaction and confidence by the respondents, as frustration was a major reason for poor data quality.

4.2 Second Research Cycle: Success

4.2.1 Action-planning

We needed a method to measure user preferences that combined the advantages of *MaxDiff* with a better usability and that fits our requirements. We identified the *Q-Sort* as the method that best matched our requirements, as it is quick and easy to use, applicable to a set of seven or more items, and solid regarding the statistical analysis of the data (Chrzan and Golovashkina, 2006).

4.2.2 Action-taking

We operationalized the method as follows: We developed the Q-Set for eight items using five columns $[-2,-1,0,+1,+2]$, with the two outermost columns having one row and the three middle columns having two rows each. Each respondent was asked to answer two *Q-Sorts* as visualized in figure 1: Nine different items need to be placed inside the pyramid-like structure having five columns. As a speciality, people are allowed to re-arrange their choices which was realized by using sticky notes that had to be glued onto their intended spot and could be detached and put on a different cell (just as the item labeled "1" in figure 1 is at first placed in the middle column and then moved to the next column on the left side).

A pre-test among twelve IS students yielded good results, especially regarding the usability of the method. In the main survey we were using a double-tracked approach. We collected the data through an online survey using *FlashQ*[2] and offline with the Android application *qResearch* [3] on a tablet computer, where students interviewed people in a pedestrian area.

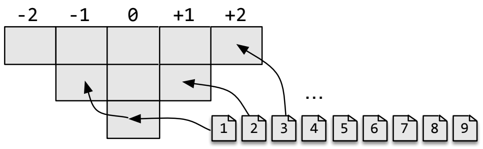

Figure 1. Exemplary Q-Sort structure.

4.2.3 Evaluating

A total of 179 people were interviewed (91 offline and 88 online), with ages from 16 to 68 years. The mean age of the online respondents (26) was slightly below that of the offline respondents (32). The majority of the respondents (129 of 179) were male. In contrast to the pre-test using *MaxDiff*, analyzing the data of the second test yielded good results, as they clearly depicted the respondents' preference for each of the two scenarios. In both scenarios (authentication and payment), a single item clearly had the highest preference, with one or two slightly lower but still high, and several others less preferred. The results confirmed our assumption that the respondents differ at their highest preferences, but show hints of indifference towards the middle, i.e. items that are neither favorited nor disliked.

4.2.4 Learning

We sought potential reasons for the success of the application of the Q-Methodology by looking at the difficulties experienced in the failure of *MaxDiff* in the first cycle. We assume that *Q-Sort* is more usable because it is less time-consuming, more consistent, and easier to use than *MaxDiff*. The time required is lower because each item has to be considered only once, rather than several times in many questions using *MaxDiff*. In addition, at the end of the *Q-Sort* process, the respondent's chosen preferences are presented graphically so he or she can check the results and make adjustments if necessary. Thus, the ease of use and the consistency are increased. What's more, respondents who used the online tool received some guidance in the sorting process and could not commit any formal mistakes, which might have contributed to the method's success.

5 Experiment

As the results gained from the second research cycle seemed promising to us, we now conduct an experiment to empirically investigate our learnings from the first and second research cycle, focusing especially on the properties of *Q-Sort* for measuring preferences. We therefore compare *Q-Sort* with other methods and identify possible circumstances, under which *Q-Sort* is a superior method for measuring user preferences.

5.1 Experiment and Setup

The experiment was carried out in two rounds a week apart. The same group of participants – students who attended a certain lecture – were asked to fill in a paper-based survey in both rounds to reflect their personal preferences with regard to the same set of payment methods using four methods: *MaxDiff*, *Ranking*, *Rating*, and *Q-Sort*. We recorded the time spent on each method by each participant. The questionnaire

[2] FlashQ Website (accessed 03/26/2015) `http://www.hackert.biz/flashq/home/`

[3] qResearch Website (accessed 03/26/2015) `http://eresearch.informatik.uni-bremen.de/`

was comprised of two parts for each method: In the first part, the actual method, e.g. *Ranking*, had to be applied and the survey part was an assessment of the characteristics of each method.

In the first round, participants were asked to evaluate each of the methods with regard to *ease of use, satisfaction, easiness* and *confidence*. *Satisfaction* measures the respondent's degree of satisfaction with the application of the method (Bruner et al., 2001). *Confidence* refers to the respondents subjective assessment of the validity of his preference choice in the way he thinks he has been able to properly reproduce his actual preferences (Petty et al., 2002). *Ease of use* measures the respondent's opinion regarding the amount of time, effort, and complexity involved in using the method (Dabholkar, 1994). *Easiness* measures the respondent's perceived ease in applying the method (Tybout et al., 2005). We use both, *easiness* and *ease of use*, as they are established constructs to measure usability and thus render a broader view of the usability of the methods.

To evaluate the four methods, participants answered a set of three to four questions for each of the named constructs on a *Likert scale* (Likert, 1932) that resulted in a total of eleven questions. A discrete analog 7-point Likert scale was chosen over a 5-point variant because of its greater sensitivity (Diefenbach et al., 1993). Each of the eleven questions was designed and formulated to cover a single scale item that contributes to one of the four constructs (see Bruner et al. (2001)). Table 2 provides insight into which questions covered which item and to which construct the items relate to. Questions Q_9 and Q_8 contribute to two constructs, *easiness* and *ease of use*, at the same time.

Affected Construct	**Question**	**1 . . .**	**Scale Items**	**. . . 7**
Satisfaction	Q_1	dissatisfied	↔	satisfied
Satisfaction	Q_6	unpleasant	↔	pleasant
Satisfaction	Q_7	frustrating	↔	contented
Confidence	Q_2	not confident at all	↔	very confident
Confidence	Q_3	not certain at all	↔	very certain
Confidence	Q_4	not valid	↔	valid
Ease of Use	Q_5	a lot of time	↔	short time
Ease of Use	Q_{11}	slow	↔	fast
Ease of Use & Easiness	Q_9	complicated	↔	simple
Ease of Use & Easiness	Q_{10}	a lot of effort	↔	little effort
Easiness	Q_8	difficult	↔	easy

Table 2. Overview of questions, including the targeted construct and corresponding items.

In the second round, in contrast to the first round, we omitted the eleven questions asked for each method. So the sole purpose of the evaluation of the methods in the second round was to assess *reproducibility* and *concordance* between the participant's individual answers by comparing the answers given in the first and second rounds.

5.2 Data Collection

For the first round, a total of 105 questionnaires were collected, from which 86 (about 83 %) were considered for the following analysis. All survey participants were between 18 and 31 years old ($mean_{age}$ = 21.6 years), with 18 of them females and 68 of them males.

For the second round, the same group of people was targeted, but only 41 of the first-round survey participants were present or could be identified unambiguously because some people were concerned about entering personal information on the questionnaire. After both surveys were completed, some questionnaires were discarded as defective with regard to validity or because they were incomplete. Following the data cleansing, we analyzed the final set of data in more detail, as discussed section 5.3.

5.3 Analysis and Interpretation

After cleansing the data of irregular and inconsistent entries, we performed the statistical evaluation using Microsoft's Excel and the *R* framework (R Development Core Team, 2008).

5.3.1 Loading Factors

Table 3 shows an overview of the individual *loading factors* that were computed using a factor analysis. The values provide evidence concerning whether the chosen scale is appropriate for measuring a specific construct using the corresponding items. For the factor analysis we made *a priori* assumptions about the number of factors based on the design of the questionnaire.

Question **Construct**	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}
MaxDiff:											
Satisfaction	0.69	–	–	–	–	0.83	0.80	–	–	–	–
Ease of Use	–	–	–	–	0.86	–	–	–	0.37	0.90	0.90
Confidence	–	0.82	0.73	0.67	–	–	–	–	–	–	–
Easiness	–	–	–	–	–	–	–	0.85	0.94	0.45	–
Q-Sort:											
Satisfaction	0.60	–	–	–	–	0.85	0.77	–	–	–	–
Ease of Use	–	–	–	–	0.88	–	–	–	0.28	0.82	0.84
Confidence	–	0.82	0.89	0.82	–	–	–	–	–	–	–
Easiness	–	–	–	–	–	–	–	0.83	0.90	0.38	–
Ranking:											
Satisfaction	0.74	–	–	–	–	0.94	0.84	–	–	–	–
Ease of Use	–	–	–	–	0.76	–	–	–	0.61	0.90	0.85
Confidence	–	0.91	0.90	0.91	–	–	–	–	–	–	–
Easiness	–	–	–	–	–	–	–	0.93	0.87	0.67	–
Rating:											
Satisfaction	0.78	–	–	–	–	0.94	0.68	–	–	–	–
Ease of Use	–	–	–	–	0.78	–	–	–	0.70	0.75	0.95
Confidence	–	0.88	0.91	0.82	–	–	–	–	–	–	–
Easiness	–	–	–	–	–	–	–	0.94	0.89	0.62	–

Table 3. Loading factors for MaxDiff, Q-Sort, Ranking and Rating.

5.3.2 Statistical Figures of Constructs

Tables 4 and 5 summarize the *standard deviation (SD)*, *variance (VAR)*, and *mean* for each construct and method. Table 4 contains information related to *confidence* and *satisfaction*, whereas table 5 provides the statistical figures for *ease of use* and *easiness*.

Method	$SD_{\text{Confidence}}$	$VAR_{\text{Confidence}}$	$mean_{\text{Confidence}}$	$SD_{\text{Satisfaction}}$	$VAR_{\text{Satisfaction}}$	$mean_{\text{Satisfaction}}$
MaxDiff	1.6690	2.7855	4.0659	1.8952	3.3986	4.0552
Q-Sort	1.4356	2.0611	4.8566	1.6533	2.7335	5.0349
Ranking	1.4687	2.1570	4.9496	1.4726	2.1686	5.4767
Rating	1.5080	2.2741	4.5891	1.5487	2.3986	5.1890

Table 4. Statistical figures: Confidence and Satisfaction.

The mean values for confidence show that the level of confidence in the results of the *MaxDiff* approach is much lower than that of the other three methods. The level of confidence in the *Rating* results is in the middle, while *Ranking* and *Q-Sort* have the best results. There are no significant differences in the calculated standard deviations (SD), but the SD is large for all four methods. Overall, it seems that the level of confidence is dependent on the complexity of the approach; using more complex approaches like *MaxDiff* lowers users' confidence in their results.

The users considered the *Ranking* method the most satisfying, while *Q-Sort* and *Rating* are in the middle, and *MaxDiff* is considered the least satisfying approach by a significant margin. The SD of *MaxDiff* is the largest, while the other methods are at intervals of < 0.2. Overall, the results for satisfaction are good, but the users were least satisfied with the *MaxDiff* method, although it's high SD shows large differences in the users' perceptions of it. To prove significance of the findings, we have applied a *two-tailed two-sample t-test* to test a *null-hypothesis*.

Method	SD_{Easiness}	VAR_{Easiness}	$mean_{\text{Easiness}}$	$SD_{\text{EaseOfUse}}$	$VAR_{\text{EaseOfUse}}$	$mean_{\text{EaseOfUse}}$
MaxDiff	1.7555	3.0817	4.6590	1.8952	3.3986	4.0552
Q-Sort	1.4828	2.1988	5.4264	1.6533	2.7335	5.0349
Ranking	1.5232	2.3202	5.3450	1.4726	2.1686	5.4767
Rating	1.5700	2.4651	5.1318	1.5487	2.3986	5.1890

Table 5. Statistical figures: Easiness and Ease of Use.

Two of the tested methods, *Q-Sort* and *Ranking*, perform best with the construct of easiness. Participants regard *Q-Sort* as the easiest method, while *Ranking* leads the other constructs. *MaxDiff* again has the lowest rating, but the gap is smaller than in the previous figures. Regarding ease of use, *Ranking* is again best, followed by *Q-Sort* and *Rating*, which are close together, and finally *MaxDiff*, which is considered to have the worst ease of use. Like the SD of satisfaction, the SD of *MaxDiff* is larger than that of the others.

Therefore, in all tested dimensions *MaxDiff* always scores worst and always has the highest SD. *Ranking* usually has the highest mean (confidence, satisfaction, and ease of use), while *Rating* and *Q-Sort* (best in easiness) are not far behind.

5.3.3 Concordance

In order to validate the preferences expressed with each method, the results from all four methods were compared using a *concordance* analysis. This procedure determines whether the individual preferences participants expressed while using the four approaches yielded the same or at least a similar result. Results that vary significantly from method to method indicate that at least one of the compared methods did not measure preferences correctly. However, as preferences are subjective it is not possible to determine which preferences are correct.

Some pre-processing is necessary in order to be able to compare the individual results, so we normalized the underlying data for each method for each participant. As this paper is focused on assessing the *Q-Sort* methodology, we investigated three relationships:

- *Q-Sort* vs. *MaxDiff*
- *Q-Sort* vs. *Ranking*
- *Q-Sort* vs. *Rating*

Once, the *similarity* between the named pairs was computed for each participant, we aggregated all results. This procedure abstracts from the individual preferences and measures whether the preferences determined using the four methods were consistent for the whole group. No severe discrepancies among the results of the four methods were found.

5.3.4 Longitudinal Concordance

As we conducted the same survey on two dates one week apart, we were able to determine whether the participants' preferences were reproducible. We relied on personal information entered in each survey to compare the two surveys that were completed by the same person. Forty-one participants completed the surveys on both dates.

The *Kendall coefficient of concordance* (Kendall, 1938) served as basis for the evaluation. Table 6 summarizes our findings with regard to longitudinal concordance of the four evaluated methodologies.

However, these preliminary results are not comparable, as correctly reproducing one's preferences using *MaxDiff* is more likely than correctly reproducing one's preferences using *Ranking*. The reason for this result is clear: Whereas a participant using *MaxDiff* is likely to have the same favorite and least favorite choices one week later, the differences between two adjacent ranks are more subtle. Similar difficulties apply when comparing preferences retrieved using *Ranking* and *Rating* (Carterette, 2009).

Method	Concording Pairs	Discording Pairs
MaxDiff	569	169
Q-Sort	236	133
Ranking	167	202

Table 6. Concordance of results from the first and second rounds.

The results show that *MaxDiff* yields very similar results for both phases of the experiment whereas *Ranking* exhibits a higher volatility in people's preferences. One should note, that reproducing one's preference is just by chance more likely to result in the same preference as for instance for *Q-Sort*, because there are only 5 discrete levels while *Ranking* has 9. Thus, in case a respondent swapped to adjacent ranks (e.g. exchanged the methods formerly in position 5 and 6) two discordant pairs result even though both preferences are very similar. In contrast, swapping two items from the same column in *Q-Sort* will not cause a discordant pair.

5.3.5 Duration

To provide reliable data on the time required to express one's preferences using the four methods, we recorded the time each participant needed to complete the task using each method.

For the given visualizations typical *Tukey boxplot* settings were used, i.e. the end of the *whiskers* demarcate the *highest* value within the 1.5 *interquartile range (IQR)* of the upper quartile and the *lowest* value within the 1.5 IQR of the lower quartile. The *vertical bar* inside a box represents the *median*, and the boxes visualize the IQR stretching from the 25th to the 75th percentile (cf. McGill et al. (1978)).

The boxplot in figure 2 shows the combined durations for each method. Outliers in the plot have been removed, leaving the following sample sizes: n_{MaxDiff}=83, $n_{\text{Q-Sort}}$=83, n_{Ranking}=82 and n_{Rating}=80. No outliers were removed for the plot in figure 3, so the original sample size of n=86 applies.

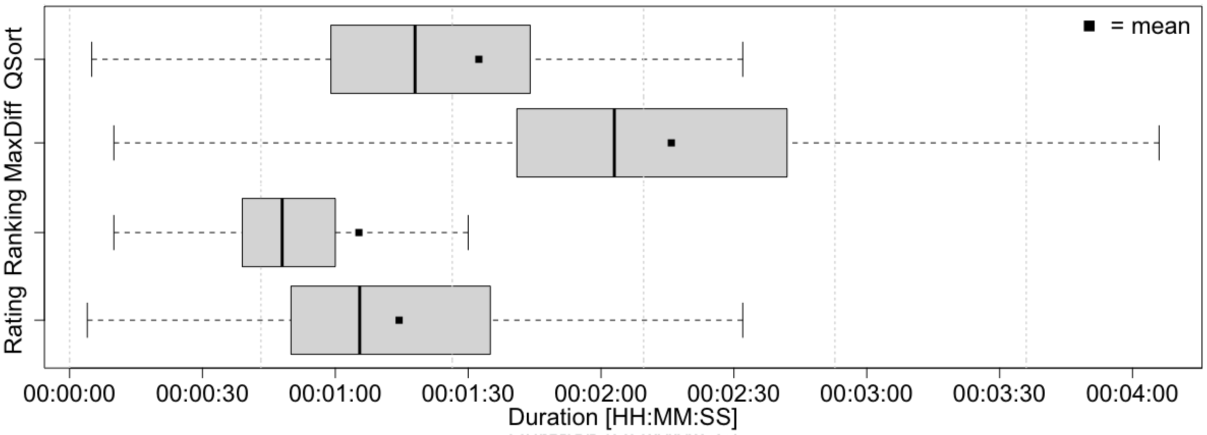

Figure 2. Boxplot of Durations.

Interestingly, when looking at the answers given to the question Q_{11} (see table 2) asking whether the subjective impression of a method was rather slow or fast, the results in figure 3 do correlate. The *MaxDiff* method was subjectively and objectively the slowest among the four methods, while *Ranking* was identified as the fastest.

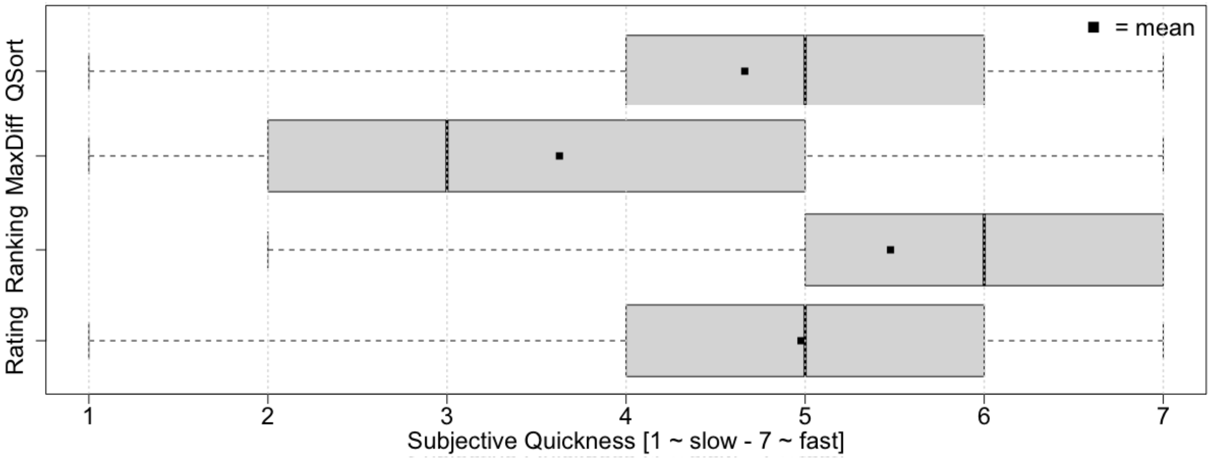

Figure 3. Boxplot of Subjective Quickness.

Figure 2 shows that some survey participants claimed to have finished the task of constructing their preferences in less than ten seconds, which is unlikely. However, introducing a threshold and filtering out these unrealistic durations would falsify the evaluation, so we kept these values but emphasize their doubtful plausibility.

6 Conclusion and Outlook

The development of IS requires that choices are made in regard to the alternative designs a system can follow. Especially in the case of a system that customers will use, it is important to identify customer preferences regarding the system's functionality. From all of the principles that suggest how to make a design decision (best practices, theories, argumentational foundation), we chose a user-centric approach in which design choices that have a direct influence on the user experience are made based on the users' preferences. However, an efficient assessment of users' preferences is a challenge both methodologically and operationally, as integrating a user closely into the development process, as in the human-centered Information Systems Design, is often too time-consuming and too expensive. On the other hand, standard techniques by which to assess user preferences can also fail or deliver inaccurate results. The user experience during preference assessment can be challenging, exhausting, and frustrating, depending on the method used, and can produce unreliable results and even discourage potential users from future use of the system to be designed.

In this article, we addressed the challenge of the effective and efficient assessment of user preferences. We demonstrated how we came to the novel application of the *Q-Sort* after a series of failed attempts to assess user preferences in a concrete IS development project. We presented the results of an experiment in which we compared the *Q-Sort* to three standard techniques for assessing user preferences: *Rating*, *Ranking*, and *Maximum Difference Scaling*. Using time effort (user performance), user satisfaction, user confidence, easiness, and ease of use, as well as the consistency within and the concordance among methods as evaluation dimensions, we showed that the *Q-Sort* is a good choice for determining an interviewee's preferences. The concordance test showed that *Q-Sort* is equally suited to the established methods to represent users' preferences. Users of the *Q-Sort* have a high level of confidence that their views are correctly represented with the methodology, consider *Q-Sort* to be the easiest method in the tested set, and attest to its ease of use. In terms of the measured and subjective duration of using the method, *Q-Sort* is close to *Ranking* and equal to *Rating*.

Our contribution is twofold. For practice, we present an alternative method for assessing user preferences in the context of user-centered IS development. We recommend using *Q-Sort* in scenarios where developers must make difficult choices, as the method needs to be easy for participants to use. In addition, *Q-Sort* can be used even when the number of alternatives is relatively large. As the number of items the brain can process at the same time is limited (Miller, 1956), methods like *Ranking* and *Rating* become more unwieldy as the number of alternatives increases. By using *Q-Sort*, users can permanently reorder the items and build smaller subsets for reordering. *Q-Sort* is also a good approach if knowing the preferences in the middle section – those that are not the most or least preferable – is not important. Especially if the decision items are similar and a definitive order is difficult to determine, *Q-Sort* unburdens the participant. If time and the subject's satisfaction are critical, *Q-Sort* is preferred over other sophisticated approaches, such as *MaxDiff*.

Second, our evaluation of competing methods for the assessment of user preferences can facilitate the scientific discourse on the choice of methods for assessing user preferences during IS development. Adapting methods from other disciplines (in this case, psychology) helps to sustain innovation in IS.

Our study is subject to a number of limitations that suggest directions for future research. While our sample size of 105 participants provides useful initial insights, to validate our findings or even reveal additional insights, future research could increase the sample size. While evaluating the results of the experiment, we had to reject some data sets when participants applied the methods incorrectly (e.g., assigning ranks twice). Future research could use IT-supported experiments to enforce the correct application of methods. Finally, this work stopped at the point at which a favorable design choice was revealed. Additional user tests could support this finding.

The assessment of user preferences during the development phase is central to the successful design and development of an information system. This paper makes a step toward a more efficient way of assessing user preferences.

Acknowledgment

This paper has been written in the context of the research project *CrowdStrom*. The project is founded by the German Federal Ministry of Education and Research (BMBF), promotion sign 01FK13019. We thank the project management agency German Aerospace Center (PT-DLR).

References

Abras, Chadia et al. (2004). "User-Centered Design." In: *Berkshire Encyclopedia of Human-Computer Interaction, Volume 2*. Ed. by William Sims Bainbridge. Great Barrington, MA: Berkshire, pp. 763–768.

Alwin, Duane F. and Jon A. Krosnick (1985). "The measurement of values in surveys: A comparison of ratings and rankings." *Public Opinion Quarterly* 49 (4), 535–552.

Bacon, Don R. (2003). "A comparison of approaches to importance-performance analysis." *International Journal of Market Research* 45 (1), 55–71.

Basili, Victor R. and H. Dieter Rombach (1988). "The TAME Project: Towards Improvement-Oriented Software Environments." *IEEE Transactions on Software Engineering* 14 (6), 758–773.

Bate, Ian and Neil Audsley (2004). "Flexible design of complex high-integrity systems using trade offs." In: *Proceedings of the 8th International Symposium on High Assurance Systems Engineering (HASE '04)*. Tampa, FL, pp. 22–31.

Becker, Jörg et al. (2013). "Designing Interaction Routines in Service Networks: A Modularity and Social Construction Based Approach." *Scandinavian Journal of Information Systems* 25 (1), 37–68.

Bijker, Wiebe E. (1987). "The social construction of Bakelite: Toward a theory of invention." In: *The Social Construction of Technological Systems*. Ed. by Wiebe E. Bijker et al. Cambridge, MA: MIT Press, pp. 159–187.

Brown, Stephen R. (1993). "A primer on Q methodology." *Operant Subjectivity* 16 (3/4), 91–138.

— (1996). "Q methodology and qualitative research." *Qualitative Health Research* 6 (4), 561–567.

Bruner, Gordon C. et al. (2001). "Marketing Scales Handbook: A Compilation of Multi-Item Measures, Vol. 3." In: American Marketing Association Chicago. Chicago, IL: American Marketing Association.

Carterette, Ben (2009). "On rank correlation and the distance between rankings." In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. Boston, MA, p. 436.

Chrzan, Keith and Natalia Golovashkina (2006). "An empirical test of six stated importance measures." *International Journal of Market Research* 48 (6), 717–740.

Clegg, Chris W. (2000). "Sociotechnical principles for system design." *Applied Ergonomics* 31 (5), 463–477.

Clements, Paul et al. (2001). "Evaluating a Software Architecture." In: *Evaluating Software Architectures: Methods and Case Studies*. Ed. by Paul Clements et al. Boston, MA: Addison-Wesley.

Cohen, Steven H. (2003). *Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation*. Tech. rep. Sequim, WA: Sawtooth Software, Inc.

Dabholkar, Pratibha A. (1994). "Incorporating choice into an attitudinal framework: analyzing models of mental comparison processes." *Journal of Consumer Research* 21 (1), 100–118.

Diefenbach, Michael A. et al. (1993). "Scales for assessing perceptions of health hazard susceptibility." *Health Education Research* 8 (2), 181–192.

Dorst, Kees and Nigel Cross (2001). "Creativity in the design process: co-evolution of problem-solution." *Design Studies* 22 (5), 425–437.

Eckert, Jochen and René Schaaf (2009). "Verfahren zur Präferenzmessung – Eine Übersicht und Beurteilung existierender und möglicher neuer Self-Explicated-Verfahren." *Journal für Betriebswirtschaft* 59 (1), 31–56.

Feather, Norman T. (1973). "The measurement of values: Effects of different assessment procedures." *Australian Journal of Psychology* 25 (3), 221–231.

Finn, Adam and Jordan J. Louviere (1992). "Determining the appropriate response to evidence of public concern: the case of food safety." *Journal of Public Policy & Marketing* 11 (1), 12–25.

Gasson, Susan (2003). "Human-Centered vs. User-Centered Approaches to Information System Design." *Journal of Information Technology Theory and Application* 5 (2), 29–46.

Green, Paul E. and V. Srinivasan (1990). "Conjoint analysis in marketing: new developments with implications for research and practice." *Journal of Marketing* 54 (4), 3–19.

Hevner, Alan R. et al. (2004). "Design science in information systems research." *MIS Quarterly* 28 (1), 75–105.

Hirschheim, Rudy et al. (1991). "Information systems development as social action: Theoretical perspective and practice." *Omega* 19 (6), 587–608.

Howcroft, Debra et al. (2004). "What me may learn from the social shaping of technology approach." In: *Social Theory and Philosophy for Information Systems*. Ed. by John Mingers and Leslie P. Willcocks. Chichester, UK: John Wiley, pp. 329–371.

IDEO (2011). *Human-Centered Design Toolkit: An Open-Source Toolkit To Inspire New Solutions in the Developing World*. 2nd ed. Bloomington, IN: Authorhouse.

Kendall, Maurice G. (1938). "A new measure of rank correlation." *Biometrika* 30 (1/2), 81–93.

Lansing, Jens et al. (2013). "Cloud Service Certifications: Measuring Consumers' Preferences for Assurances." In: *Proceedings of the 21st European Conference on Information Systems (ECIS '13)*. Utrecht, Netherlands.

Likert, Rensis (1932). "A technique for the measurement of attitudes." *Archives of Psychology* 22 (140), 1–55.

Louviere, Jordan J. (1991). *Best-worst scaling: A model for the largest difference judgments*. Working Paper. Alberta, Canada: University of Alberta.

Louviere, Jordan J. et al. (1994). "Retail Research Methods." In: *Handbook of Marketing Research*. Ed. by H. J. Houston. 2nd ed. New York, NY: McGraw-Hill.

Lyytinen, Kalle et al. (2008). "A framework to build process theories of anticipatory information and communication technology (ICT) standardizing." *International Journal of IT Standards and Standardization Research* 6 (1), 1–38.

Malhotra, Yogesh and Dennis F. Galletta (2004). "Building systems that users want to use." *Communications of the ACM* 47 (12), 88–94.

Matzner, Martin et al. (2015). "Crowdsourcing-Ladedienste durch Kleinanbieter als innovatives Geschäftsmodell (CrowdStrom)." In: *Dienstleistungsinnovationen für Elektromobilität: Märkte, Geschäftsmodelle, Kooperationen*. Ed. by Daniel Beverungen et al. Stuttgart, Germany: Fraunhofer-Verlag, pp. 129–142.

McGill, Robert et al. (1978). "Variations of box plots." *The American Statistician* 32 (1), 12–16.

Miller, George A. (1956). "The magical number seven, plus or minus two: some limits on our capacity for processing information." *Psychological Review* 63 (2), 81–97.

Müller, Florian H. and Elisabeth Kals (2004). "Die Q-Methode. Ein innovatives Verfahren zur Erhebung subjektiver Einstellungen und Meinungen." *Forum Qualitative Sozialforschung* 5 (2), 1–17.

Munson, J. Michael and Shelby H. McIntyre (1979). "Developing practical procedures for the measurement of personal values in cross-cultural marketing." *Journal of Marketing Research* 16 (1), 48–52.

Peffers, Ken et al. (2006). "The design science research process: a model for producing and presenting information systems research." In: *Proceedings of the 1st Conference on Design Science Research in Information Systems and Technology (DESRIST '06)*. Claremont, CA, pp. 83–106.

Petty, Richard E. et al. (2002). "Thought confidence as a determinant of persuasion: the self-validation hypothesis." *Journal of Personality and Social Psychology* 82 (5), 722–741.

Pohl, Klaus (2010). *Requirements Engineering: Fundamentals, Principles, and Techniques*. 1st ed. Springer.

Pottie, Gregory J. (1995). "System design choices in personal communications." *IEEE Personal Communications* 2 (5), 50–67.

Prieto-Diaz, Rubén and Peter Freeman (1987). "Classifying software for reusability." *IEEE Software* 4 (1), 6–16.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rokeach, Milton (1973). *The nature of human values*. New York, NY: Free Press.

Sattler, Hendrik and Susanne Hensel-Börner (2003). "A comparison of conjoint measurement with self-explicated approaches." In: *Conjoint Measurement*. Ed. by Andreas Gustaffson et al. Berlin / Heidelberg, Germany: Springer, pp. 147–159.

Segars, Albert H. and Varun Grover (1998). "Strategic information systems planning success: an investigation of the construct and its measurement." *MIS Quarterly* 22 (2), 139–163.

Simon, Herbert A. (1996). *The Sciences of the Artificial*. 3rd ed. Cambridge, MA: MIT Press.

Srinivasan, V and Chan Su Park (1997). "Surprising robustness of the self-explicated approach to customer preference structure measurement." *Journal of Marketing Research* 34 (2), 286–291.

Stephenson, William (1935). "Correlating persons instead of tests." *Journal of Personality* 4 (1), 17–24.

— (1936a). "A new application of correlation to averages." *British Journal of Educational Psychology*.

— (1936b). "The inverted factor technique." *British Journal of Psychology. General Section* 26 (4), 344–361.

— (1953). *The Study of Behaviour: Q-technique and its Methodology*. Chicago, IL: University of Chicago Press.

Thomas, Dominic M. and Richard T. Watson (2002). "Q-sorting and MIS research: A primer." *Communications of the Association for Information Systems* 8, 141–157.

Thurstone, Louis L. (1927). "A law of comparative judgment." *Psychological Review* 34 (4), 273–286.

Tybout, Alice M. et al. (2005). "Information accessibility as a moderator of judgments: The role of content versus retrieval ease." *Journal of Consumer Research* 32 (1), 76–85.