How to Quantify User Experience: Fuzzy Comprehensive Evaluation Model Based on Summative Usability Testing

Ronggang Zhou

Department of Industrial Engineering, Tsinghua University,
Beijing 100084, China
zhourg@tsinghua.edu.cn

Abstract. The concept of usability is complicated and fuzziness. Fuzzy theory is developed to provide comprehensive evaluation capabilities in the presence of imprecise and uncertain information. Starting with the ISO 9241 dimensions (effectiveness, efficiency and satisfaction), a fuzzy comprehensive model based on fuzzy theory for evaluating usability is proposed instead of conventional methods. The model has ability to assess user experience comprehensively with defuzzied score. Combined with data of summative usability, it can be applied to benchmark product usability, and a case study indicated the approach can quantify user experience directly and comprehensively.

Keywords: user experience, usability, usability testing, fuzzy comprehensive evaluation, analytic hierarchy process (AHP).

1 Introduction

As technology advance, usability has become an important criterion for decision making for end-users and consumers to chose, and users are less willing to put up with uncomfortable product when there are many competitive alternatives. So product usability captures more devotion from product designers and developers for their competitive purpose in the market. Nowadays, usability has become a special field consisted of multi-disciplines, called usability engineering. And many useful methods are available to evaluate the usability. However, less effective tools or approaches are efficient to evaluate comprehensively product's usability integrating objective and objective measure, since usability is a concept of fuzzy and its definition is dependent. The purpose of this study is to propose a comprehensive evaluation model based on fuzzy evaluation approach. The model can be used to measure the level of products usability in usability engineering processes, such as for designers and developers to know the best one in the corresponding stage of develop.

1.1 Definitions of Usability

As a core term in human-computer interaction, usability has been defined by many researchers in many ways [1] [2] [3] [4] [5] [6]. By focusing on the perception of the

N. Aykin (Ed.): Usability and Internationalization, Part II, HCII 2007, LNCS 4560, pp. 564–573, 2007.

product, Shackel proposed an operational definition of usability, and provided a set of usability criteria [1]. They were effectiveness, learnability, flexibility, and attitude or satisfaction. The definition has been generally accepted in usability community [7]. Another well-accepted definition of usability which received more attention from HCI was offered by Nielsen, he described five operational usability dimensions: learnability, memorability, efficiency, satisfaction and errors [2]. Based on the effort of whole usability community, the international Standards Organization (ISO) attempted to establish standards definition on usability, and defined usability as "the extent to which a product can be used by specified users to achieve specified goals with effective, efficiency and satisfaction in a specified context of use" [5]. However the dimensions of usability have been described by ISO/IEC 9126-1 as understandability, learnability, operability, and attractiveness [6]. From the overview of the usability definition, usability could be a combination of different dimensions, such as effectiveness, usefulness, learnability, flexibility, attitude/likeable, memorability, efficiency, satisfaction, errors, understandability, operability and attractiveness. So in some degree "the concept usability is ill defined in research and practice alike. Usability can mean different things to different people, even when it is defined, it still remains intuitive, circular, or elusive." [8], and the meaning of usability is context dependent and still ambiguous [1] [9] [10].

1.2 Attempts to Evaluate Usability Comprehensively

The definition of usability is related with usability measurement, "what we mean by the term usability is to a large extent determined by how we measure it" [4]. Many different metrics can be used for measuring one dimension of usability. For example, with binary task completion, accuracy, error rate, recall and/or completeness, we can measure the effective [10]. In a summative usability evaluation, several metrics are available to the analyst for benchmarking the usability of a product for comparing with its previous versions or competitor's systems. But generally it is difficult to make a comparison between different evaluations, since metrics, test tasks and numbers of task are used differently. If analyst could draw a comprehensive evaluation for overall usability, the comparison would become possible. From the literatures, we have seen some attempts to derive a single measure based on data of usability evaluation.

Only based on objective data of user performance time, key stroke time and error rate, Babiker et al derived a metric for measuring overall usability of hypertext system [11]. The metric was based on three individual important attribute: access and navigation, orientation, and user interaction. They found their metric correlated to subjective assessment measures, but it could not be generalized to other systems since proper weights need to be determined. With the method of Principle Components Analysis, Sauro et al also tried to derive a way to represent system or task usability in a single, standardized and summated metric, and they claimed that the metric do include all usability aspects, such as effectiveness, efficiency and satisfaction[12]. But the evaluated aspects were weighted equally.

Focusing on user's personal interactive experience with a product, several well-known subjective usability questionnaires were developed such as Software Usability Measurement Inventory (SUMI) [13], the Questionnaire for User Interaction (QUIS)

[14] [15], and Post-Study system Usability Questionnaire (PSSUQ) [16] [17]. The authors of these questionnaires do not necessarily intend for the questionnaires to act as a single measure of usability [12]. Based on human information processing theory and eight human factors considerations which are relevant to software usability, Purdue Usability Testing Questionnaire (PUTQ) was developed as a checklist for comparing the relative usability of different software systems [7]. Also in development of usability questionnaires for electronic mobile (MPUQ), Ryu tried to use decision making methods based on the Analytic Hierarchy Process (AHP) and linear regression analysis to make comprehensive usability evaluation of mobile [18]. These questionnaires can provide a subjective assessment for recently completed tasks and there were claimed to derive a reliable and low-cost standardized measure of the overall usability or quality of use of a system, but they are only suitable to subjective assessment and are not appropriate for integrating objective data.

These methods are not enough dynamic to apply to the practice for evaluating overall usability of product. Since the complication and fuzziness of the usability, the selection of evaluation approach is very important. Fuzzy theory is developed to provide decision-making capabilities in the presence of imprecise and uncertain information. Starting with the ISO 9241 dimensions (effectiveness, efficiency and satisfaction), this paper aims to propose a comprehensive evaluation model with the approach of fuzzy comprehensive evaluation integrating the AHP, and apply it to present a single usability score based on summative usability testing.

2 A Proposed Usability Comprehensive Evaluation Model

In this section, first we provided general description for fuzzy comprehensive evaluation and how to use the AHP to weight the evaluated factors. Then we proposed a comprehensive evaluation model for usability.

2.1 General Description of Fuzzy Comprehensive Evaluation

Fuzzy analytic hierarchy evaluation is the process of evaluating an objective utilizing the fuzzy set theory. When evaluating an objective, multiple related factors must be considered comprehensively in order to give an appropriate, non-contradicting and logically consistent judgment. The general steps of fuzzy evaluation may be simplified as the following [19]:

Step 1: Determining a set of evaluation factors. With these factors we can get a structural index system for evaluation. Assuming that the objective being evaluated contains n factors, then the index set can be represented as $U=\{u_1, u_2, \ldots, u_n\}$.

Step 2: Determining a set of appraisal grades. The appraisal set can be represented as $V=\{v_1, v_2, \ldots, v_m\}$, for instance {*excellent, good, medium, poor, very poor*} could be used as appraisal comment for specific objective.

Step 3: Setting fuzzy matrix for general evaluation. In this case, we'll get the mapping from U to V. For a specific factor, the appraisal is $R_i=\{r_{i1}, v_{i2}, \ldots, v_{im}\}$. The overall fuzzy appraisal matrix of all n factors can be mapped a fuzzy relationship:

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{n2} & r_{n2} & \cdots & r_{nm} \end{bmatrix}. \tag{1}$$

Step 4: Determining the weight of evaluation factors. In making a comprehensive evaluation, the importance of each factor should be quantified. The weight vector can be represented by A $(a_1, a_2, ..., a_n)$, which can be formulated by the AHP.

Step 5: Getting the appraisal result. The overall appraisal result set of comprehensive evaluation is B, presented as follows.

$$B=(b_1, b_2, b_3, ..., b_m) = A \circ R. \tag{2}$$

Where, b_j could be operated by many operation models, such as M ($\wedge$, $\vee$), M ($\cdot$, $\vee$) and M ($\cdot$, $\oplus$) [20]. In this study, every single factor should be considered. So the M ($\cdot$, $\oplus$) was used for calculated b_j, where " $\oplus$ " defined as α+β=min (1, α+β), then the model is

$$b_j = \sum_{i=1}^{n} a_i r_{ij} = \min\left\{1, \ \sum_{i=1}^{n} a_i r_{ij}\right\}. \tag{3}$$

2.2 How to Determine the Weight Vector by the AHP

In this paper, we used the AHP to obtain the weight vector A. The procedures may be simplified as follows [21][22]:

Step 1: Based on pair-comparison of n factors shown in Table 1, the weight comparison could be represented in n×n matrix as follows:

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}. \tag{4}$$

Each a_{ij} of the matrix represents the importance intensity of factor A_i over factor A_j. The a_{ij} value is supposed to be an approximation of the relative importance of A_i to A_j, i.e., a_{ij} =(W_i/W_j). Each of a_{ij} $(i,j=1,2,\ldots,n)$ follows $a_{ji}=1/ a_{ij}$, for $a_{ij}\neq 0$.

Table 1. lineal scale of preferences in the pair-wise comparison process

Numerical rating	Judgments of preferences between factor *i* and factor *j*.
1	factor *i* is equally important to factor *j*
3	factor *i* is slightly more important than factor *j*
5	factor *i* is clearly more important than *j*
7	factor *i* is strongly more important than factor *j*
9	factor *i* is extremely more important than factor *j*
2, 4, 6, 8	Intermediate values

Step 2: Calculating the weight vector *A*. We can use the method of ANC (average of normalized columns) to estimate the vectors of weights function, ANC can be presented as:

$$w_i = \frac{1}{n}\sum_{j=1}^{n}\frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}. \tag{5}$$

Step 3: Computing consistence ratio of the judgments matrix. Accordingly, Saaty defined the consistency ratio as:

$$\mathrm{CR}=CI\,/\,RI. \tag{6}$$

The CR is a measure of how a given matrix compares to a purely random matrix in terms of their consistency indices. A value of CR≤0.1 is considered acceptable. RI is the average random index, which is a statistical value. For a 3 ×3 matrix, the value of RI is 0.58. And where consistency index (CI) was defined as:

$$\mathrm{CI}=(\lambda_{\max}-n)/(n-1), \tag{7}$$

where n is the number of factors, and $\lambda_{\max}$ represents the maximum eigenvalue of the pairwise comparison matrix, the closer the $\lambda_{\max}$ is to n the more consistent, and the $\lambda_{\max}$ can be formulated by:

$$\lambda_{\max} = \sum_{i=1}^{n}\frac{(A\vec{w})_i}{nw_i}. \tag{8}$$

2.3 A Weighted Hierarchical Index Proposed for Evaluating Usability

Usability cannot be directly measured, but we can construct it into attributes that can be measured. The choice of such attributes not only fleshes out what usability means, it also raises the question if that which is measured is a valid indicator of usability [10]. The framework of usability provided by ISO is pervasive [5], and was selected as a basis for structure of usability evaluation index in this paper like previous studies [12] [18] [23], i.e. effectiveness, efficiency and satisfaction structured as three attributes of usability. Since the two measures are product-independent and used most frequently, in order to structure a universal usability evaluation index for different systems, *task success* and *task completion time* were selected as a single metric for measuring effectiveness and efficiency respectively. The PSSUQ was developed exclusively for measurement of satisfaction for user testing, and had the highest percentages of redundancy with the other sets of questionnaire items [18], and so was chosen for measuring user's satisfaction after a test in this study.

Since single metric was employed to measure effectiveness, efficiency and satisfaction, we only need to determine the weight vector of the three attributes for overall usability. We had a six-expert panel to perform pair-wise comparison according to Table 1. They discussed together and gave agreeable pair-wise comparison with respect to the three attributes of usability, and they would repeat the process if the CR>0.1. Table 2 presented the matrix of pair-wise comparisons, and according to section 2.2, and the weight vector *A* could be given as (0.443, 0.169, 0.387).

Table 2. Pairwise comparison with respect to user satisfaction

	Effective	Efficiency	Satisfaction	Weight
Effective	1	3	1	0.443
Efficiency	1/3	1	1/2	0.170
Satisfaction	1	2	1	0.387

Note: λmax=3.018, CI=0.009, CR=0.016.

So the fuzzy evaluation model can be constructed as shown in Fig. 1.

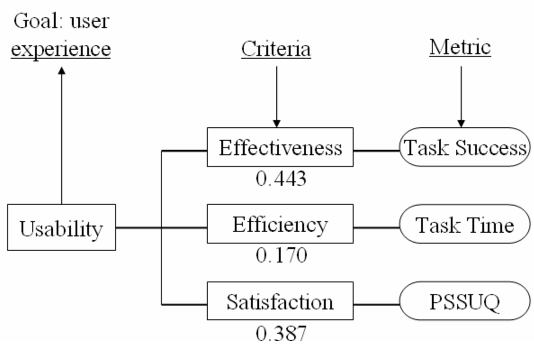

Fig. 1. Fuzzy comprehensive usability evaluation model

2.4 Determining the Fuzzy Member Function for Appraisal Matrix *R*

In this study, the metric of task success was valued by 0~1, “0” indicates one participant can not finish a test task, “1” means he complete the test task very well, intermediate value means corresponding degree of success. Satisfaction was scaled by PSSUQ, which is 7-point scale. Task time was recorded from the beginning to the end of task. How to value or record the success and time was described operationally. When determining the membership function for factors, corresponding score of each task on each metric would be ranked as “*excellent, good, medium, poor* or *very poor*”. Table 3 presented the membership mapping, which was determined based on expert’s experience.

Table 3. The membership mapping for metric score ranking

Ranking	*very poor*	*poor*	*medium*	*good*	*excellent*
Success	$0 \le x < 0.3$	$0.3 \le x < 0.6$	$0.6 \le x < 0.8$	$0.8 \le x < 0.95$	$0.95 \le x \le 1$
Time	$x < 0.3$	$0.3 \le x' < 0.6$	$0.6 \le x' < 0.8$	$0.8 \le x' < 0.95$	$0.95 \le x'$
Satisfac.	$1 \le x < 2$	$2 \le x < 3.5$	$3.5 \le x < 5.5$	$5.5 \le x < 6.5$	$6.5 \le x \le 7$

In Table 3 *x* is the mean performance. Before test, the shortest complete time was given as a expect value for each task, so *x*’ was transformed by *x* according to the following formula:

$$x' = 1 - \frac{x - E}{E} = 2 - \frac{x}{E}, \tag{9}$$

where "*E*" means expectable shortest task time, when $x'=1$, the performance on task time is the best. Then the factors in fuzzy relation matrix could be calculated as following formula [24]:

$$R_{ij} = \text{(Num. of corresponding average rank) / (Num. of the participants)} \qquad (10)$$

3 Case Study

In order to illustrate fuzzy comprehensive evaluation model could be applied to benchmark usability of product, one summative usability testing process was used as an example in this section. Based on integrated user-centered design approach [25], a software product was developed. Before releasing, a standard summative usability testing was conducted in a standard usability testing lab to benchmark the usability of the product [2] [25]. There were 16 typical users participated the testing. And the testing was processed by one experienced facilitator, and two usability engineers collected the data respectively as observers in the watching room.

3.1 Determining the Fuzzy Appraisal Matrix

According to section 2.1 and 2.4, the average performance of all tasks on each metric is calculated. Then each of mean value is ranked as "excellent, good, medium, poor or very poor", which is presented judgment set in the paper. According to Eq. (10), the fuzzy appraisal matrix for these three factors was obtained. The process could be illustrated from Table 4, which indicated the membership for task success.

Table 4. The membership mapping for task success value ranking

	M_{success}	*Excellent*	*Good*	*Medium*	*Poor*	*Very Poor*
P1	0.945		×			
P2	0.963	×				
P3	0.981	×				
P4	0.985	×				
P5	0.955	×				
P6	0.966	×				
P7	0.949		×			
P8	0.963	×				
P9	0.946		×			
P10	0.935		×			
P11	0.956	×				
P12	0.952	×				
P13	0.955	×				
P14	0.964	×				
P15	0.937		×			
P16	0.989	×				
	Total	11	5	0	0	0
	R_j	0.6875	0.3125	0	0	0

Similar way, we can get membership mapping for task time and satisfaction. So the fuzzy appraisal matrix could be presented as following.

$$R = \begin{bmatrix} 0.6875 & 0.3125 & 0 & 0 & 0 \\ 0.25 & 0.3125 & 0.0625 & 0.25 & 0.125 \\ 0.125 & 0.625 & 0.25 & 0 & 0 \end{bmatrix} \cdot \tag{11}$$

3.2 Getting the Appraisal Result

In this paper, we consider very single factor overall, so *B* was calculated based on Eq. (2).

$$\begin{aligned} B = A \circ R = (0.4434, \quad 0.1692, \quad 0.3874) \circ \begin{bmatrix} 0.6875 & 0.3125 & 0 & 0 & 0 \\ 0.25 & 0.3125 & 0.0625 & 0.25 & 0.125 \\ 0.125 & 0.625 & 0.25 & 0 & 0 \end{bmatrix} . \\ = (0.3956, \quad 0.3711, \quad 0.1758, \quad 0.0156, \quad 0.0313) \end{aligned} \tag{12}$$

This was the final appraisal vector, and it can be defuzzified to a comprehensive score [25]. In this paper, we defined excellent, good, medium, poor, very poor in appraisal grading as 95, 82, 67, 50, 31, respectively, so the appraisal vector *B* can be defuzzified according to the following formula:

$$a = {\sum_{i-1}^{m} b_i^2 a_i} \Big/ {\sum_{i-1}^{m} b_i^2}, \tag{13}$$

where a is the defuzzified score, a_1=95, a_2=82, a_3=67, a_4=50, a_5=31, b_i is appraisal vector. Base on the appraisal vector, the defuzzified score was 86.63, which can present the comprehensive usability of the software.

4 Discussion and Conclusion

Based on the fuzzy evaluation theory, a model for evaluating usability of a system was proposed instead of conventional methods. Fuzzy comprehensive evaluation theory is an effective approach for quantifying and qualifying the uncertain, and is appropriate to evaluate usability comprehensively. Based on the fuzzy evaluation model, the defuzzified score can provide a synthetic judgment for user experience of product using. Integrated with data of summative usability testing (e.g. performance measurements), the model can be used to measure the level of products usability in corresponding developed processes, such as for designers and developers to know the best one in the stage of the competitive analysis process or to validate the success of their own new product before releasing, since it can provide one continuous variable that can be used for hypothesis testing statistically. In addition, the approach can also be applied to structure other usability evaluation data systematically.

References

1. Shackel, B.: Usability - context, framework, design and evaluation. In: Shackel, B., Richardson, S. (eds.) Human factors for informatics usability. Cambridge, pp. 21–38 (1991)
2. Nielsen, J.: Usability Engineering. Academic Press, San Diego (1993)
3. Rubin, J.: Handbook of Usability Testing: How to plan, design and conduct effective tests. John Wiley & Sons, New York (1994)
4. Barnum, C.M.: Usability Testing and Research. Longman Publications, New York (2002)
5. ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) – part 11: Guidance on usability. International Organization for Standardization (1998)
6. ISO/IEO 9126 -1: Software engineering- product quality – part 1: Quality model. International Organization for Standardization (2001)
7. Lin, H.X., Choong, Y.-Y., Salvendy, G.: A Proposed Index of Usability: A Method for Comparing the Relative Usability of Different Software Systems. Behaviour & Information Technology 16, 267–278 (1997)
8. Kim, K.: A model of digital library information seeking process (DLISP model) as a frame for classifying usability problems. Doctoral dissertation, The State University of New Jersey. New Brunswick, New Jersey (2002)
9. Newman, W., Taylor, A.: Towards a methodology employing critical parameters to deliver performance improvements in interactive systems. In: Proceedings of IFIP TC.13 International Conference on Human-Computer Interaction, pp. 605–612. IOS Press, Amsterdam (1999)
10. Hornbæk, K.: Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies 64, 79–102 (2006)
11. Babiker, E.M., Fujihara, H., Boyle, C.D.B.: A metric for hypertext usability. In: Proc. 11th Annual International Conference on Systems documentation, pp. 95–104. ACM Press, New York (1991)
12. Sauro, J., Kindlund, E.: A Method to Standardize Usability Metrics Into a Single Score. In: CHI 2005, Portland, OR, pp. 401–409. ACM Press, New York (2005)
13. Kirakowski, J.: The Software Usability Measurement Inventory: Background and usage. In: Jordan, P., Thomas, B., Weerdmeester, B. (eds.) Usability Evaluation in Industry, pp. 169–178. Taylor and Francis, London (1996)
14. Chin, J.P., Diehl, V.A., Norman, K.L.: Development of an instrument measuring user satisfaction of the human-computer interface. In: Proceedings of SIGCHI ’88. ACM/SIGCHI, New York pp. 213–218 (1988)
15. Harper, B.D., Norman, K.L.: Improving User Satisfaction: The Questionnaire for User Interaction Satisfaction Version 5.5. In: Proceedings of the 1st Annual Mid-Atlantic Human Factors Conference. Virginia Beach, VA, pp. 224–228 (1993)
16. Lewis, J.R.: IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human–Computer Interaction 7, 57–78 (1995)
17. Lewis, J.R.: Psychometric evaluation of the PSSUQ using data from five years of usability studies. International Journal of Human-Computer Interaction 14, 463–488 (2002)
18. Ryu, Y.S.: Development of Usability Questionnaires for Electronic Mobile Products and Decision Making Methods. Doctoral dissertation, Virginia Polytechnic Institute and State University. Blacksburg, Virginia (2005)

19. Liang, Z., Yang, K., Sun, Y., Yuan, J., Zhang, H., Zhang, Z.: Decision support for choice optimal power generation projects: Fuzzy comprehensive evaluation model based on the electricity market. Energy Policy 34, 3359–3364 (2006)
20. Lan, H., Ding, Y., Hong, J.: Decision support system for rapid prototyping process selection through integration of fuzzy synthetic evaluation and an expert system. International Journal of Production Research 43, 169–194 (2005)
21. Saaty, T.L.: The analytic hierarchy process. McGraw Hill, New York (1980)
22. Hsiao, S-W., Chou, J-R.: A Gestalt-like perceptual measure for home page design using a fuzzy entropy approach. International Journal of Human-Computer Studies 64, 137–156 (2006)
23. Park, K.S., Lim, C.H.: A structured methodology for comparative evaluation of user interface designs using usability criteria and measures. International Journal of Industrial Ergonomics 23, 379–389 (1999)
24. Vredenburg, K., Isensee, S., Righi, C.: User-Centered Design: An Integrated Approach. Prentice Hall, New Jersey (2001)
25. Kuo, Y.-F., Chen, L.-S.: Using the fuzzy synthetic decision approach to assess the performance of university teachers in Taiwan. International journal of management 19, 593–604 (2002)