Information & Management 46 (2009) 397–403

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im

The measurement of user satisfaction with question answering systems

Chorng-Shyong Ong [a], Min-Yuh Day [a,b,*], Wen-Lian Hsu [b]

[a] Department of Information Management, National Taiwan University, No. 50, Lane 144, Sec. 4, Keelung Road, Taipei 106, Taiwan
[b] Institute of Information Science, Academia Sinica, 128 Academia Road, Sec. 2, Nankang, Taipei 115, Taiwan

ARTICLE INFO

Article history:
Received 29 September 2007
Received in revised form 18 June 2009
Accepted 26 July 2009
Available online 15 August 2009

Keywords:
Information system success
Measurement
Question answering
QAS
User satisfaction

ABSTRACT

Question Answering Systems (QAS) are receiving increasing attention from IS researchers, particularly those in the information retrieval and natural language processing communities. Evaluation of an IS's success and user satisfaction are important issues, especially for emerging online service systems using the Internet. Although many QAS have been implemented, little work has been done on the development of an evaluation model for them. Our purpose was to develop a validated instrument to measure user satisfaction with QAS (USQAS). The proposed validated instrument was intended as a reference for the design of QAS from a user's perspective.

1. Introduction

Evaluation of the success of an IS and user satisfaction with it are important issues in the field of information management [7–10], especially for online service systems using the Internet. Basically, evaluation models are used to understand users' needs and identify important dimensions and factors in the development of systems in order to broaden their acceptance. With the rapid growth in recent years, QAS have emerged as important applications. Hence, they have received attention from IS researchers, particularly those in the information retrieval and natural language processing communities [5,6].

The purpose of our study was to develop an evaluation model for QAS based on a review of IS models and theories. Thus, the new model incorporates constructs from both the user satisfaction and technology acceptance literature. Since we focus on the user's perspective, we believe that the proposed model could facilitate the design of QAS and thereby enhance user satisfaction and acceptance.

A QAS is a special type of information retrieval system that allows users to input questions in natural language and retrieve answers from a collection of documents.

Kokubu et al. [4] addressed the link between user satisfaction and the performance of QAS. More specifically, they investigated the relationship between the rank of a correct answer and the Proportion of Satisfied Users (PSU), where PSU is defined as the number of users that were satisfied with a given list of answer candidates for a question, divided by the total number of users. To improve user satisfaction, Kokobu et al. suggested that QA system developers should set a goal in terms of the distribution of correct answers over ranks, instead of a single Mean Reciprocal Rank (MRR) value.

Although many QAS have been implemented, little work has been done on the development of an evaluation model for them. Appropriate evaluation would motivate research by providing suggestions for the overall improvement of the architecture and behavior of QAS. Such models should provide feedback on a system's architecture and the impact of its behavior on the user, thereby facilitating improvements in the system. Evaluation models could also help to determine the extent to which a particular system meets its requirements and demonstrate its research value.

Most evaluation models focus on system-centered evaluation; user-centered evaluation has attracted less attention. However, if we are to build a practical QAS, we must achieve a performance level that satisfies the majority of users. Therefore we proposed an evaluation model of successful QAS from the user's perspective. Our goal was to answer two questions.

- How do individual users evaluate the success of a QAS? and
- What factors influence an individual user's evaluation of QAS success?

[*] Corresponding author at: Department of Information Management, National Taiwan University, No. 50, Lane 144, Sec. 4, Keelung Road, Taipei 106, Taiwan. Tel.: +886 2 27883799x1366; fax: +886 2 26518660.
E-mail address: myday@iis.sinica.edu.tw (M.-Y. Day).

0378-7206/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.im.2009.07.004

2. Theoretical foundations

2.1. Domain of user satisfaction with QAS

A QAS is a special type of information retrieval system that can retrieve answers from a collection of documents (such as the World Wide Web or a local collection) to questions input in natural language. It is generally agreed that a QAS requires more complex natural language processing (NLP) than other types of information retrieval.

An evaluation model should provide feedback about a system's architecture as well as its impact on the user, thereby facilitate improvements in the system. Such a model could also help to determine the extent to which a particular system meets requirements and demonstrate its value.

Existing user satisfaction instruments in the IS field are considered inappropriate for QAS for the following reasons. (1) QAS that use natural language processing and information retrieval techniques are distinct from those employed in the end-user computing or traditional data processing environments. (2) Recognizing the changes in the IT environment, Doll and Torkzadeh have argued that the existing user satisfaction instruments are inappropriate for the EUC environment in which end-users develop and/or interact directly with specific applications, like QAS.

Venkatesh et al. noted that investigating user acceptance of a new technology is one of the most mature research areas in contemporary information systems (IS) literature. Such research has resulted in the development of several theoretical models, with roots in information systems, psychology, and sociology, which explain over 40% of the variance in individual intentions to use a particular technology. Confronted with a choice of a multitude of models, researchers find that they must "pick and choose" constructs across the models, or choose a "favored model" and largely ignore the contributions of alternative models. Thus, there is a need for a review and synthesis of the literature pertaining to different models with the objective of developing a unified view of user acceptance.

Information systems researchers have long studied how and why individuals adopt new information technologies. There are several streams of IS research, one of which focuses on explaining individual acceptance of a technology by considering intention to use or actual usage as a dependent variable. Other streams focus on the success of implementations at the organizational level, or on task-technology fit.

Wixom and Todd [11] suggested that research on perceptions of IS success can be categorized into two primary research streams—user satisfaction literature and technology acceptance literature. However, as these two approaches have developed in parallel, they have not been reconciled or integrated.

Venkatesh et al. reviewed extant user acceptance models to assess the state of knowledge with respect to understanding individual acceptance of new information technologies. Their review identified eight prominent models and discussed their similarities and differences. They then developed the Unified Theory of Acceptance and Use of Technology (UTAUT), the goal of which is to understand usage as a dependent variable.

To improve our understanding of the evaluation of systems, we reviewed five major IS theories:

1. Fishbein and Ajzen's Theory of Reasoned Action (TRA), drawn from social psychology, it has been used to predict a wide range of behaviors. Its core constructs are *attitude toward behavior* (an individual's positive or negative feelings about performing the target behavior) and *subjective norm* (the person's perception that most people who are important to him think he should or should not perform the behavior in question).
2. The Technology Acceptance Model (TAM), an adaptation of TRA, is designed specifically for modeling and predicting information technology acceptance and usage in the work environment. Unlike TRA, TAM excludes the attitude construct in order to better explain the *intention to use* variable. It was extended to include the concept of subjective norm as an added predictor of intentions in mandatory use settings. Its core constructs are *perceived usefulness* (the degree to which a person believes that using a particular system would enhance his or her job performance) and *perceived ease of use*. (The degree to which a person believes that using a particular system would be free of effort.)
3. Ajzen's Theory of Planned Behavior (TPB) incorporated a third determinant of behavioral intention, *perceived behavioral control* (a person's motivation is influenced by the perceived difficulty of the behavior), determined by control beliefs and perceived power. Positive or negative perceptions might reflect past experience, anticipation of upcoming circumstances, and the attitudes associated with the influential norms that surround the individual.
4. Venkatesh et al.'s Theory of Acceptance and Use of Technology (UTAUT); accounted for dynamic influences like organizational context, user experience, and demographic characteristics by incorporating four key moderators (gender, age, voluntariness, and experience). They also theorized that four constructs (performance expectancy, effort expectancy (the degree of ease associated with the use of the system), social influence (the degree to which an individual perceives that important others believe he or she should use the new system), and facilitating conditions (the degree to which an individual believes that an organizational and technical infrastructure exists to support use of the system)) were direct determinants of user acceptance and use behavior. Performance expectancy (the degree to which an individual believes that using the system will help him or her to obtain gains in job performance) was assumed to depend on perceived usefulness, extrinsic motivation, job-fit, relative advantage, and outcome expectations.
5. Wixom and Todd's Theoretical Integration of User Satisfaction and Technology Acceptance (TIUSTA) model [11] was constructed by developing an integrated research model that distinguished object-based beliefs and attitudes about a system from beliefs and attitudes about using the system. Then, they proposed a theoretical logic that linked user satisfaction and technology acceptance models. TIUSTA thus built a bridge between design and implementation decisions and system characteristics (a core strength of user satisfaction), as well as between decisions about and predictions of usage (a core strength of technology acceptance). As such, TIUSTA constituted an important step toward providing conceptual clarity about these two critically important streams of research. They further argued that there was a need to develop a refined understanding of the relationships proposed in TIUSTA.

2.2. Conceptualization of user satisfaction with QAS (USQAS)

We propose an evaluation model for QAS based on models of IS user satisfaction and technology acceptance. The fundamental concept was inspired by the TRA believe–attitude–intention–behavior theory.

We adopted three dimensions of quality from the new DeLone and McLean Information Systems (IS) Success Model: information quality, systems quality, and service quality. Satisfaction also has three dimensions: information satisfaction, systems satisfaction, and service satisfaction. The quality of the information provided by a QAS is shaped by four dimensions: completeness, accuracy, format, and currency. *Completeness* represents the degree to which the system provides all necessary information; *accuracy* represents the user's perception that the information is correct; *format* represents the user's perception of how well the information is presented; and *currency* represents the user's perception of the degree to which the information is up-to-date.

Information quality (IQ) is a measure of the quality of the content of an IS, but data quality (DQ) is often used as a synonym for it. Wang and Strong [12] developed a hierarchical framework that captured the aspects of data quality important to data consumers. They collected 118 data quality attributes and consolidated them into twenty dimensions, which were then grouped into four categories. The resulting framework had four DQ categories:

- intrinsic, consisting of accuracy, objectivity, believability, and reputation;
- contextual, consisting of value-added, relevancy, timeliness, completeness, and appropriate quantity;
- representational, consisting of interpretability, ease of understanding, representational consistency, and concise representation; and
- accessibility, consisting of accessibility and access security.

Meanwhile, *system quality*, is measured on five dimensions: reliability, flexibility, integration, accessibility, and timeliness. *Reliability* measures the dependability of the system's operation; *flexibility* involves the way the system adapts to the changing demands of the user; *integration* refers to the way the system merges data from various sources; *accessibility* is the ease with which information can be accessed or extracted from the system; and *timeliness* assumes timely responses to requests for information or action.

*Service qualit*y, the user's judgment of the overall excellence of a QAS, depends on assurance, the level of certainty a user has of the quality of the service provided, empathy, the degree to which a service employee shows understanding and sympathizes with a user's situation, and responsiveness, the reaction time of the service.

Information satisfaction is the extent to which an individual's attitude influences the gap between *expectations* and *perceived performance of the information provided*. Similarly, *system satisfaction* is the extent to which an individual's attitude influences the gap between *expectations* and the *perceived performance of the system*; while *service satisfaction* is the extent to which an individual's attitude influences the gap between *expectations* and the *perceived performance of the servic*e.

Beliefs about quality tend to affect satisfaction. Information satisfaction and system satisfaction shape beliefs about *perceived usefulness* and *perceived ease of use*, respectively. *Perceived usefulness, perceived ease of use,* and *service quality* tend to shape an individual's attitude towards a QAS and his/her intention to use it. *Intention to use* in turn shapes *individual usage* of the QAS. Many researchers consider that an individual's intention to use a system is significantly correlated to his/her actual usage, which in turn is an indicator of acceptance of an IS. In summary, user satisfaction can be regarded as a function of perceived ease of use and perceived usefulness, and is likely to lead to IS acceptance and success.

3. Research methodology

3.1. Generation of scale items

Although a number of items could be used to measure the USQAS construct, it was necessary to define its theoretical meaning and conceptual domain so that we could develop appropriate measures and obtain valid results. We defined user satisfaction with a QAS as a user's overall evaluation of the system. We selected 35 items based on prior research, including user information satisfaction, end-user computing satisfaction, TAM, and relevant QAS-related articles. These 35 items represented the nine dimensions underlying the USQAS construct. They were used to form the initial pool of items for the USQAS scale. To ensure that we did not omit any important attributes, we conducted three QAS-related focus group interviews with two professors, five doctoral students, and ten practitioners. As a result, we were able to refine the items and eliminate unnecessary content. Specifically, 10 items were deleted because of ambiguity or redundancy, and two items were added. After careful examination of the interview results, we compiled a 27-item list that constituted our domain for USQAS measurement. Pre-testing and pilot testing of the measures were conducted by selected users from the QAS field, as well as by experts in the area. Only three ambiguous items were modified in this stage.

To obtain a quick overall measure of satisfaction prior to detailed analysis, the items had to represent the concept about which generalizations were to be made in order to ensure the validity of the scales' content. Five global items adapted from previous inventories were used to evaluate the criterion-related validity and nomological validity of the USQAS instrument. Two items for measuring overall satisfaction were taken from Doll and Torkzadeh. Specifically, "Are you satisfied with the system?" was altered to "As a whole, I am satisfied with the QAS." and "Is the system successful?" was changed to "As a whole, the QAS is successful". *Behavioral intention to use* was evaluated by two items taken from Venkatesh and Davis: "Assuming I had access to a QAS, I intend to use it", and "Given that I had access to a QAS, I predict that I would use it". The following item for measuring favorable post-usage behavior (recommending the system to others), was adapted from Devaraj et al. [2]: "I will recommend the QAS to others." Hence, our initial USQAS instrument consisted of 27 items, including the five global items; it used a seven-point Likert scale, with anchors ranging from "strongly disagree" to "strongly agree". The global measures were used to analyze the criterion-related validity of the instrument, and to measure the overall satisfaction with the QAS prior to detailed analysis. In addition to the USQAS measuring items, the questionnaire contained demographic questions. For each question, respondents were asked to circle the response that best described their level of agreement. All the items, including initial and global items, were modified to make them relevant to the QAS context. Appendix A presents the items used in our study.

3.2. Sample and procedure

The data for the USQAS instrument was collected from 276 users of an Internet QAS (the Academia Sinica QAS, ASQA). The respondents self-administrated their 27-item questionnaire. For each question, respondents were asked to circle the response that best described their level of agreement with the statements. Of the 276 surveys, 235 useful responses were returned; a response rate of 85%. All the respondents had prior experience in using QAS. Most were students (29.4%) and engineers (23.4%), and 69% were male. Their average age was 31.5 years. Forty-nine percent held a university degree; and 43% held graduate degrees.

4. Scale purification

Since the primary purpose of this study was to develop a reliable and accurate general instrument capable of measuring USQAS, pooling the sample data from Internet users was considered appropriate.

Several tests were conducted to refine the initial 27 items (excluding the five global items). Reliability tests suggested that screening the data would improve reliability levels. First, we calculated the reliability coefficients of the scales using Cronbach's alpha. It seemed appropriate to assume that USQAS was a simple construct before using exploratory factor analysis to identify its

Table 1
Factor analysis results: principal component extraction.

Item code	Original item code	Factor			
		Ease of use	Usefulness	Service quality	Information quality
Ease of use					
E1	Q27	**0.91**	0.19	0.07	0.10
E2	Q22	**0.89**	0.11	0.04	0.16
E3	Q25	**0.89**	0.09	0.13	0.10
E4	Q24	**0.88**	0.07	0.10	0.15
E5	Q26	**0.86**	0.10	0.02	0.15
Usefulness					
U1	Q20	0.09	**0.83**	0.17	0.25
U2	Q18	0.12	**0.79**	0.21	0.29
U3	Q16	0.10	**0.78**	0.33	0.25
U4	Q19	0.05	**0.77**	0.17	0.31
U5	Q21	0.16	**0.76**	0.16	0.14
Service quality					
S1	Q12	0.09	0.19	**0.86**	0.18
S2	Q13	0.04	0.15	**0.86**	0.20
S3	Q11	0.11	0.24	**0.86**	0.18
S4	Q14	0.07	0.29	**0.82**	0.17
Information quality					
I1	Q2	0.14	0.38	0.19	**0.79**
I2	Q4	0.11	0.26	0.21	**0.76**
I3	Q1	0.28	0.37	0.26	**0.74**
I4	Q3	0.27	0.27	0.21	**0.73**
Cronbach's alpha		0.94	0.91	0.92	0.89
Eigenvalue		8.07	3.29	1.73	1.04
Cumulative variance explained (%)		44.85	63.11	72.74	78.54

Items with a factor loading greater than 0.5 are shown in bold.

underlying dimensions. Based on this assumption, we found that the reliability of the initial 27 items was 0.94.

For the remaining sets of items, item-to-total correlations were examined to eliminate irrelevant content. We screened the data to identify items that showed very low item-to-total correlations (i.e., <0.5). Because the minimum value of the item-to-total correlation was above 0.5, no items were deleted in the stage.

Exploratory factor analysis was conducted to purify the instrument by eliminating items that did not load on an appropriate high-level construct. The analysis identified the underlying factors or the dimensional composition of the USQAS instrument. The 235 responses were examined using principal component factor analysis as the extraction technique, with varimax rotation.

To improve the convergent and discriminant validity of the instrument through exploratory factor analysis, four widely used decision rules were applied to identify the factors underlying the USQAS construct:

(1) a minimum eigenvalue of 1 was taken as a cut-off value for extraction;
(2) items with a factor loading of less than 0.5 on all factors, or greater than 0.5 on two or more factors were deleted;
(3) a simple factor structure was assumed; and
(4) for the sake of parsimony, single-item factors were excluded.

The factor analysis and item deletion process was repeated until all items had been analyzed. As a result, we obtained a 4-factor, 18-item instrument. The results confirmed the existence of four factors with eigenvalues greater than 1, which cumulatively accounted for 78.5% of the total variance, as shown in Table 1. There were no items with cross-factor loadings above 0.5. The significant loading of all the items on a single factor indicated unidimensionality, and the fact that no cross-loadings of items were found supported the discriminant validity of the instrument.

5. Assessment of reliability and validity

5.1. Reliability

Reliability can be determined by using Cronbach's alpha to assess the internal consistency of the items representing each factor. The 18-item instrument had a high reliability (0.92) far exceeding the minimum standard suggested for basic research. Furthermore, the minimum value of each corrected item-to-total correlation was above 0.5 (minimum = 0.52), suggesting that the instrument had good reliability (see Table 2).

5.2. Content validity

The content validity refers to the representativeness of the item content domain: the manner in which the questionnaire and its items are built to ensure the reasonableness of the claims of content validity. The rigorous procedures used to select the USQAS constructs to form the initial items, personal interviews with experts, and the iterative procedures of scale purification imply that the USQAS instrument has strong content validity.

5.3. Criterion-related validity

Criterion-related validity is the effectiveness of a measure in predicting behavior in specific situations. It is determined by comparing the correlation coefficient test scores with the external criterion or overall satisfaction. In our study, we determined the correlation between the total scores of the USQAS instrument (the sum of 18 items) and the measures of criterion validity (the sum of five global items used to measure overall satisfaction with QAS). The results showed that the 18-item USQAS instrument had a criterion-related validity of 0.62 and a significance level of 0.000, suggesting acceptable criterion-related validity.

Table 2
Corrected item-to-total correlations.

Item code	Original item code	Item description	Corrected item-to-total correlation
E1	Q27	My interaction with the QAS is clear and understandable.	0.56
E2	Q22	Learning to use the QAS is easy.	0.56
E3	Q25	It is easy for me to be come skillful at using the QAS.	0.56
E4	Q24	I find it easy to use the QAS to do what I want it to do.	0.55
E5	Q26	I find the QAS easy to use.	0.52
U1	Q20	Using the QAS would enhance my effectiveness on the job.	0.64
U2	Q18	I would find the QAS useful in my job.	0.68
U3	Q16	Using the QAS would improve my job performance.	0.70
U4	Q19	Using the QAS in my job would increase my productivity.	0.62
U5	Q21	Using the QAS would make it easier to do my job.	0.58
S1	Q12	The QAS is dependable.	0.59
S2	Q13	The QAS employees provide prompt service to users.	0.55
S3	Q11	The QAS has up-to-date hardware and software.	0.63
S4	Q14	The QAS employees have the knowledge to do their job well.	0.61
I1	Q2	Information provided in the QAS is easy to understand.	0.69
I2	Q4	Information provided in the QAS is relevant.	0.60
I3	Q1	Information provided by the QAS is complete.	0.77
I4	Q3	Information provided in the QAS is personalized.	0.67

Table 3
Correlation matrix of measures.

	Ease of use					Usefulness					Service quality				Information quality			
	E1	E2	E3	E4	E5	U1	U2	U3	U4	U5	S1	S2	S3	S4	I1	I2	I3	I4
E1	**1.00**																	
E2	**0.81**	**1.00**																
E3	**0.80**	**0.80**	**1.00**															
E4	**0.79**	**0.78**	**0.78**	**1.00**														
E5	**0.79**	**0.74**	**0.72**	**0.73**	**1.00**													
U1	0.22	0.25	0.21	0.18	0.19	**1.00**												
U2	0.25	0.24	0.21	0.23	0.23	**0.75**	**1.00**											
U3	0.23	0.20	0.21	0.23	0.23	**0.72**	**0.74**	**1.00**										
U4	0.17	0.20	0.19	0.17	0.17	**0.71**	**0.67**	**0.68**	**1.00**									
U5	0.25	0.24	0.25	0.23	0.22	**0.61**	**0.60**	**0.65**	**0.57**	**1.00**								
S1	0.18	0.16	0.22	0.19	0.17	0.35	0.38	0.48	0.36	0.34	**1.00**							
S2	0.13	0.14	0.17	0.16	0.10	0.34	0.36	0.45	0.34	0.29	**0.74**	**1.00**						
S3	0.21	0.18	0.23	0.23	0.16	0.39	0.46	0.51	0.39	0.37	**0.76**	**0.77**	**1.00**					
S4	0.18	0.16	0.22	0.18	0.12	0.44	0.46	0.52	0.42	0.35	**0.75**	**0.70**	**0.75**	**1.00**				
I1	0.27	0.29	0.29	0.28	0.27	0.53	0.54	0.56	0.60	0.46	0.41	0.35	0.43	0.42	**1.00**			
I2	0.22	0.26	0.23	0.27	0.25	0.46	0.50	0.47	0.42	0.42	0.39	0.37	0.38	0.41	**0.68**	**1.00**		
I3	0.40	0.38	0.38	0.40	0.40	0.55	0.58	0.57	0.57	0.48	0.45	0.45	0.48	0.45	**0.80**	**0.65**	**1.00**	
I4	0.35	0.41	0.33	0.37	0.36	0.48	0.52	0.50	0.47	0.37	0.37	0.40	0.40	0.38	**0.66**	**0.57**	**0.72**	**1.00**

The within-factor correlations are shown in bold.

5.4. *Construct validity*

The construct validity can be demonstrated by validating the theory behind the instrument. Researchers have used various validation strategies to establish it, including item-to-total correlations, factor analysis, and assessment of convergent and discriminant validity, which demonstrates construct validity by showing that an instrument not only correlates with variables with which it should correlate, but also does not correlate with variables from which it should differ.

We used a correlation matrix approach. Convergent validity determines whether associations between scales of the same factor are higher than zero and large enough to proceed with the discriminant validity test. Table 3 presents the correlation matrix of the measures. The smallest within-factor correlations were: *ease of use* = 0.72; *usefulness* = 0.57; *service quality* = 0.70; and *information quality* = 0.57. These correlations are significantly higher that zero ($p < 0.000$) and large enough to proceed with discriminant tests.

Discriminant validity is determined by counting the number of times an item correlates more with items of other factors than with items of its own factor. For example, the lowest within-factor correlation for usefulness was 0.57, however, one of the correlations of content with items of other factors was larger than 0.57, i.e., the number of violations was 1. For discriminant validity, the count should be less than 50% of the potential comparisons. Table 3 shows only four violations for potential comparisons, suggesting adequate discriminant validity. Hence, the observed convergent and discriminant validity suggest the adequacy of the measurements used in our study.

6. Conclusion

We rigorously tested our proposed USQAS instrument and found that it provided a high degree of confidence in the reliability and validity of the scales. A comprehensive model for measuring USQAS was presented in Fig. 1. In our study, we developed a 4-factor, 18-item instrument for measuring USQAS. The four primary dimensions of USQAS are ease of use, usefulness, service quality, and information quality.

The measure assumes that users will be dissatisfied with a system if it does not provide information in a satisfactory form. The technical quality of a system is irrelevant, since a technically superior system would not be considered successful if it did not meet users' needs. This explains why system quality does not play a role in measuring user satisfaction with a QAS here and the absence of system quality from the final model is not surprising.

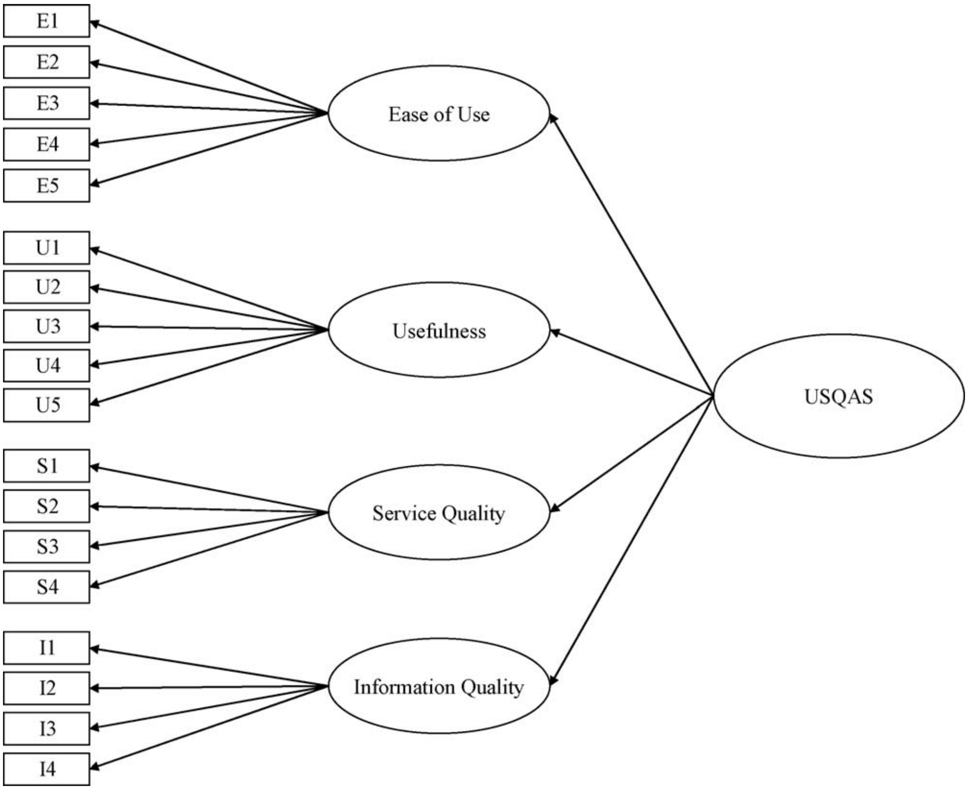

Fig. 1. A comprehensive model for measuring user satisfaction with QAS (USQAS).

Our study provides a framework for describing the primary dimensions of user satisfaction with QAS. Also, the framework can be translated into a validated instrument for measuring user satisfaction levels. A variety of statistical tests were used to demonstrate the reliability and validity of the questionnaire.

We believe the proposed evaluation model provides a framework for the design of QAS from the user's perspective and that it could help increase user acceptance of QAS.

Our study has important implications for managers and the design of efficient and effective QAS because the instrument provides a common framework for comparative analysis. For the research community, this study established a generalized instrument for USQAS.

Our study has some limitations. First, the external validity limitations should be considered when interpreting the results. We collected data from individuals who used one special QAS. This could limit the generalization of our findings to other populations or QAS.

Second, although we conducted exploratory factor analysis, a much larger sample is required to ensure greater precision. Such analysis would validate the existing underlying domain structure, and may be used to refine the structure further.

Third, since responses to the questionnaire were voluntary, they were inevitably subject to self-selection bias because users who were interested in, had used, or were currently using QAS were more likely to respond.

Finally, since this study was conducted with a single snapshot approach, the test–retest reliability could not be evaluated.

Acknowledgements

We would like to thank the editor-in-chief, Prof. Edgar H. Sibley, and the anonymous reviewers for many valuable comments and suggestions, which have enabled us to improve our paper significantly. This research was supported in part by the National Science Council of Taiwan under Grants NSC 95-2416-H-002-047 and NSC 95-2752-E-001-001-PAE.

Appendix A. The initial measurement of USQAS: Questionnaire Scales

No.	Items	References
Q1	Information provided by the QAS is complete.	[11]
Q2	Information provided in the QAS is easy to understand.	[11]
Q3	Information provided in the QAS is personalized.	[11]
Q4	Information provided in the QAS is relevant.	[11]
Q5	Information provided in the QAS is secure.	[11]
Q6	The system provided in the QAS is reliable.	[11]
Q7	The system provided in the QAS is flexible.	[11]
Q8	The system provided in the QAS is integrated.	[11]
Q9	The system provided in the QAS is accessible.	[11]
Q10	The system provided in the QAS is timely.	[11]
Q11	The QAS has up-to-date hardware and software.	[3]
Q12	The QAS is dependable.	[3]
Q13	The QAS employees provide prompt service to users.	[3]
Q14	The QAS employees have the knowledge to do their job well.	[3]
Q15	The QAS has the user's best interests at heart.	[3]

Appendix A (*Continued*)

No.	Items	References
Q16	Using the QAS would improve my job performance.	[1]
Q17	Using the QAS in my job would enable me to accomplish tasks more quickly.	[1]
Q18	I would find the QAS useful in my job.	[1]
Q19	Using the QAS in my job would increase my productivity.	[1]
Q20	Using the QAS would enhance my effectiveness on the job.	[1]
Q21	Using the QAS would make it easier to do my job.	[1]
Q22	Learning to use the QAS is easy.	[1]
Q23	I find the QAS is flexible to interact with.	[1]
Q24	I find it easy to use the QAS to do what I want to do.	[1]
Q25	It is easy for me to be come skillful at using the QAS.	[1]
Q26	I find the QAS easy to use.	[1]
Q27	My interaction with the QAS is clear and understandable.	[1]
Q28	Assuming I had access to a QAS, I intend to use it.[a]	
Q29	Given that I had access to QAS, I predict that I would use it.[a]	
Q30	I will recommend the QAS to others.[a]	[2]
Q31	As a whole, I am satisfied with the QAS.[a]	
Q32	As a whole, the QAS is successful.[a]	

[a] Criterion items.

References

[1] S. Petter, E.R. McLean, A meta-analytic assessment of the DeLone and McLean IS success model: An examination of IS success at the individual level, Information & Management 46 (3), 2009, pp. 159–166.

[2] S. Devaraj, M. Fan, R. Kohli, Antecedents of B2C channel satisfaction and preference: validating e-commerce metrics, Information Systems Research 13 (3), 2002, pp. 316–333.

[3] W.J. Doll, X.D. Deng, T.S. Raghunathan, G. Torkzadeh, W.D. Xia, The meaning and measurement of user satisfaction: a multigroup invariance analysis of the end-user computing satisfaction instrument, Journal of Management Information Systems 21 (1), 2004, pp. 227–262.

[4] T. Kokubu, T. Sakai, Y. Saito, H. Tsutsui, T. Manabe, M. Koyama, et al., The relationship between answer ranking and user satisfaction in a question answering system, in: Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-5), Tokyo, Japan, 2005.

[5] J. Lin, User simulations for evaluating answers to question series, Information Processing and Management 43 (3), 2007, pp. 717–729.

[6] J. Lin, D. Demner-Fushman, Methods for automatically evaluating answers to complex questions, Information Retrieval 9 (5), 2006, pp. 565–587.

[7] R. McHaney, R. Hightower, J. Pearson, A validation of the end-user computing satisfaction instrument in Taiwan, Information & Management 39 (6), 2002, pp. 503–511.

[8] C.S. Ong, J.Y. Lai, Measuring user satisfaction with knowledge management systems: scale development, purification, and initial test, Computers in Human Behavior 23 (3), 2007, pp. 1329–1346.

[9] C.S. Ong, J.Y. Lai, Y.S. Wang, Factors affecting engineers' acceptance of asynchronous e-learning systems in high-tech companies, Information & Management 41 (6), 2004, pp. 795–804.

[10] Y.S. Wang, Assessment of learner satisfaction with asynchronous electronic learning systems, Information & Management 41 (1), 2003, pp. 75–86.

[11] B.H. Wixom, P.A. Todd, A theoretical integration of user satisfaction and technology acceptance, Information Systems Research 16 (1), 2005, pp. 85–102.

[12] R.Y. Wang, D.M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of Management Information Systems 12 (4), 1996, pp. 5–34.

Chorng-Shyong Ong is a professor of Information Management at National Taiwan University (NTU), Taiwan. He holds a Master's degree in management science and policy studies at TSUKUBA University in Japan. He received his Ph.D. in business administration from NTU. His research interests include IS service quality, web-based services, electronic commerce and strategic management of e-business. He has published papers in *Information & Management, Computers in Human Behavior, Applied Mathematics and Computation, Pattern Recognition Letters, Lecture Notes in Computer Science (LNCS), Mathematical and Computer Modelling, Journal of Information Management, Journal of Quality,* and other journals.

Min-Yuh Day is a doctoral candidate in the Department of Information Management at National Taiwan University, Taiwan. He is also a member of the Intelligent Agent Systems Lab in the Institute of Information Science, Academia Sinica, Taiwan. His current research interests include Knowledge Management, Electronic Commerce, Information Systems Evaluation, Question Answering Systems, Data Mining and Text Mining.

Wen-Lian Hsu is a professor and distinguished research fellow in the Institute of Information Science at Academia Sinica Taipei, Taiwan, ROC. He received a B.S. from the Department of Mathematics, National Taiwan University in 1973, an M.S. and a Ph.D. in operations research from Cornell University in 1978 and 1980, respectively. In 1980, he joined Northwestern University as an assistant professor and was promoted to tenured associate professor in 1986. He joined the Institute of Information Science as a research fellow in 1989. His areas of research include graph algorithm, artificial intelligence, and bioinformatics. He developed the famous Chinese input system, GOING, which has more than 1.5 million users in Taiwan. His Chinese question answering system has won the first place twice in the International NTCIR contest. He has received many awards including the Distinguished Researcher from the National Science Council of Taiwan and the Teco Technology award. He is an IEEE Fellow.