# Development of an Evaluation Model for Question Answering Systems

Chorng-Shyong Ong [1], Min-Yuh Day [1, 2], and Wen-Lian Hsu [2], *Fellow, IEEE*
[1] *Department of Information Management, National Taiwan University, Taiwan*
[2] *Institute of Information Science, Academia Sinica, Taiwan*
*ongcs@im.ntu.edu.tw; {myday, hsu}@iis.sinica.edu.tw*

## Abstract

*Question Answering Systems (QAS) are receiving increasing attention from information systems researchers, particularly those in the information retrieval and natural language processing communities. Evaluation of an information system's success and user satisfaction are important issues in information systems research, especially for emerging online service systems on the Internet. Although many QAS have been implemented, little work has been done on the development of an evaluation model for them. The purpose of this study is to develop an evaluation model for question answering systems from user's perspective. We have established the theoretical foundation and conceptualization of the constructs for user satisfaction with question answering systems. We have developed a general instrument capable of reliably and accurately measuring user satisfaction in the context of question answering systems. The proposed evaluation model provides a framework for the design of question answering systems from the user's perspective to enhance user satisfaction and acceptance of QAS.*

**Keywords:** Evaluation Model; Measurement; Question answering system; User satisfaction

## 1. Introduction

Evaluation of an information system's success and user satisfaction are important research issues in the field of information management, especially for online service systems on the Internet. 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 of the Internet and database technologies in recent years, question answering systems (QAS) have emerged as important applications. Hence, they have received a great deal of attention from information systems researchers, particularly those in the information retrieval and natural language processing communities [32-34, 39, 43].

The purpose of this study is to develop an integrated evaluation model for question answering systems based on a literature review of information system models and theories. Based on the user's perspective, we propose an integrated reference model for the design of QAS to enhance user satisfaction and acceptance.

A great deal of research has addressed the issue of usability in the field of information retrieval. However, to the best of our knowledge, the relationship between QA performance and user satisfaction has not been measured directly. Lin et al. [26] investigated the size of supporting passages that users prefer. Allan et al.[2] examined the relationship between the quality of an information retrieval system and its effectiveness for the user, and suggested that researchers should focus on improving the performance of IR systems for difficult topics.

Although many QAS have been implemented, little work has been done on the development of an evaluation model for them [6, 22, 28, 29, 32-34, 39]. Appropriate evaluation would motivate research by providing suggestions for the overall improvement of the architecture and behavior of QAS [27]. Such models should provide feedback about a system's architecture and the impact of its behavior on the user, and thereby facilitate improvements in the system. Evaluation models could also help determine the extent to which a particular system meets certain requirements and demonstrate its research value.

As most evaluation models focus on system-centered evaluation, user-centered evaluation has attracted little attention. However, if we are to build a practical question answering system, we must achieve a performance level that satisfies the majority of users. Therefore, in this paper, we propose an integrated theoretical evaluation model of successful QAS from the user's perspective. Our goal is to answer two questions. (1) How do individual users evaluate the success of question answering systems? (2) What factors influence an individual user's evaluation of a QAS success?

The remainder of this paper is organized as follows. In Section 2, we establish the theoretical foundation and

conceptualization of the constructs for user satisfaction with question answering systems (USQAS). Section 3 describes the research methods used to generate scale items and the data collection procedure. In Section 4, we present the results of purifying the scale and identify the factor structure of the scale. In Section 5, we assess the reliability and validity of the proposed model by examining the evidence of reliability, content validity, criterion-related validity, and construct validity - including the convergent validity and the discriminant validity. Finally, in Section 6, we present our conclusions, discuss the managerial implications of the present study, and consider future research avenues.

## 2. Research Background

Satisfaction is an important construct in IS literature [1, 3, 10, 13, 14, 17, 18, 20, 31, 38, 41, 42]. As Bailey and Pearson [3] noted, satisfaction in a given situation is the sum of one's feelings or attitudes toward a variety of factors affecting that situation. In the cited work, they identified 39 factors that affect computer user satisfaction (CUS). Ives et al. [20], based on Bailey and Pearson' findings, developed a 39-item instrument for measuring user information satisfaction (UIS). Baroudi and Orlikowski [4] developed a UIS short-form questionnaire comprised of 13 scales (with 2 items per scale), which can be decomposed into three factors: "EDP staff and services", "information product", and "knowledge and involvement". Doll and Torkzadeh [13] suggested a 12-item instrument that measures the following five components of end-user computing satisfaction (EUCS): content, accuracy, format, ease of use, and timeliness. The EUCS instrument is a synthesis of the Ives et al. [20] measure of UIS, which is a widely used and well-validated instrument [13, 14, 31]. User satisfaction is defined as the extent to which users believe the information system available to them meets their information requirements [20].

In this paper, we propose a theoretical evaluation model for question answering systems based on a review and synthesis of existing models of IS user satisfaction and technology acceptance. The fundamental concept of the proposed model is inspired by the TRA believe-attitude-intention-behavior theory, which is one of the most influential theories of human behavior [16].

We adopt three dimensions of quality from the updated DeLone and McLean Information Systems (IS) Success Model[10], namely, information quality, systems quality, and service quality. Like quality, satisfaction also has three dimensions: information satisfaction, systems satisfaction, and service satisfaction. Clearly, quality will affect subsequent satisfaction. Information quality, defined as 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.

Meanwhile, system quality, which is defined as the quality of the system provided by a QAS, is shaped by five dimensions: reliability, flexibility, integration, accessibility, and timeliness. Reliability refers to the dependability of the system's operation; flexibility refers the way the system adapts to the changing demands of the user; integration refers to the way the system integrates data from various sources; accessibility refers to the ease with which information can be accessed or extracted from the system; and timeliness refers to how well the system provides timely responses to requests for information or action.

Service quality, defined as the user's judgment about the overall excellence or superiority of a QAS, is shaped by three dimensions: assurance, empathy, and responsiveness.

User satisfaction is defined as the extent to which users believe the information system available to them meets their information requirements [20]. Information satisfaction is defined as the extent to which an individual's attitude influences the gap between expectations and the perceived performance of the information provided. Similarly, system satisfaction is defined as the extent to which an individual's attitude influences the gap between expectations and the perceived performance of the system; while service satisfaction is defined as the extent to which an individual's attitude influences the gap between expectations and the perceived performance of the service.

Beliefs about information quality, system quality, and service quality tend to shape attitudes about information satisfaction, system satisfaction, and service quality, respectively. 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 the use behavior of the QAS.

## 3. Research Methodology

Although a number of potential items can be used to measure the USQAS construct, it is necessary to define the construct's theoretical meaning and conceptual domain so that we can develop appropriate measures and obtain valid results [7, 31]. We define user satisfaction with a question

answering system as a user's overall evaluation of the question answering system. The definition is important to delimit the USQAS domain and identify relevant literature from which researchers can generate sample items for an instrument. Based on prior research on instrument development, including user information satisfaction[20], end-user computing satisfaction[13-15], the technology acceptance model[9], the theoretical integration of user satisfaction and technology acceptance model[41], as well as ten relevant QAS related articles (e.g., [12, 21-23, 25, 27, 28, 36, 39, 43]), there were 35 items (excluding five global items) representing the 9 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 or items, we conducted three QAS-related focus group interviews with two professors, five doctoral students, and ten practitioners. As a result, we are able to refine the items and eliminate unnecessary content. In this stage, 10 items were deleted because of ambiguity or redundancy, and two new items were added. After careful examination of the interview results, we compiled a 27-item list, which constituted a complete 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 must represent the concept about which generalizations are to be made 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 & Torkzadeh [13]. Specifically, "Are you satisfied with the system?" was refined to "As a whole, I am satisfied with the question answering system." and "Is the system successful?" was refined to "As a whole, the question answering system is successful". Behavioral intention to use was evaluated by two items taken from Venkatesh & Davis [37], namely, "Assuming I had access to a question answering system, I intend to use it", and "Given that I had access to a question answering system, I predict that I would use it" .The following item for measuring favorable post usage behavior, i.e., recommending the system to other people, was adapted from Devaraj et al.[11]: "I will recommend the question answering system to others." Hence, an initial USQAS instrument comprised of 27 items, including the five global items, was developed using a seven-point Likert scale, with anchors ranging from "strongly disagree" to "strongly agree". The global measures can be used to analyze the criterion-related validity of the instrument, and to measure the overall satisfaction with the question answering system 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 items and global items, were modified to make them relevant to the QAS usage context.

The data used to test the USQAS instrument was obtained from 276 users of an Internet QAS (i.e., the Academia Sinica Question Answering System, ASQA). The respondents self-administered the 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 usable 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; a further 43% held graduate degrees.

## 4. Scale Purification

Since the primary purpose of this study was to develop a general instrument capable of reliably and accurately measuring USQAS in the context of question answering systems, 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 according to Churchill's [7] recommendations would improve reliability levels. First, we calculated the reliability coefficients of the scales using Cronbach's alpha [8]. It seemed appropriate to assume that USQAS was a simple construct before using exploratory factor analysis to identify its underlying dimensions. Based on this assumption, we found that the initial 27 items had a reliability of 0.941.

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 [7, 13, 20]. 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, and varimax as the rotation

**Table 1.** Corrected Item-to-total correlations

| Item Code | Original Item code | Item Description | Corrected Item-Total Correlation |
|---|---|---|---|
| E1 | Q27 | My interaction with the QAS is clear and understandable. | 0.557 |
| E2 | Q22 | Learning to use the QAS is easy. | 0.556 |
| E3 | Q25 | It is easy for me to become skillful at using the QAS. | 0.556 |
| E4 | Q24 | I find it easy to use the QAS to do what I want it to do. | 0.550 |
| E5 | Q26 | I find the QAS easy to use. | 0.517 |
| U1 | Q20 | Using the QAS would enhance my effectiveness on the job. | 0.643 |
| U2 | Q18 | I would find the QAS useful in my job. | 0.680 |
| U3 | Q16 | Using the QAS would improve my job performance. | 0.702 |
| U4 | Q19 | Using the QAS in my job would increase my productivity. | 0.618 |
| U5 | Q21 | Using the QAS would make it easier to do my job. | 0.584 |
| S1 | Q12 | The QAS is dependable. | 0.587 |
| S2 | Q13 | The QAS employees provide prompt service to users. | 0.547 |
| S3 | Q11 | The QAS has up-to-date hardware and software. | 0.628 |
| S4 | Q14 | The QAS employees have the knowledge to do their job well. | 0.608 |
| C1 | Q2 | Information provided in the QAS is easy to of understand. | 0.685 |
| C2 | Q4 | Information provided in the QAS is relevant. | 0.603 |
| C3 | Q1 | Information provided by the QAS is complete. | 0.772 |
| C4 | Q3 | Information provided in the QAS is personalized. | 0.672 |

method.

To improve the convergent validity and discriminant validity of the instrument through exploratory factor analysis, the following four widely used decision rules [19, 35] 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; 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 confirm the existence of four factors with eigenvalues greater than 1, which cumulatively account for 78.5% of the total variance. The four factors are ease of use, usefulness, service quality, and information quality. Note that there were no items with cross-factor loadings above 0.5. The significant loading of all the items on a single factor indicates unidimensionality, and the fact that no cross-loadings of items were found supports the discriminant validity of the instrument.

## 5. Assessment of reliability and validity

Reliability refers to the stability of scores over a variety of conditions [24] and can be determined by using Cronbach's alpha to assess the internal consistency of the items representing each factor. The 18-tem instrument has a high reliability of 0.92, far exceeding the minimum standard of 0.80 suggested for basic research. The reliability of each factor is as follows: ease of use = 0.94; usefulness = 0.90; service quality = 0.92; and content quality = 0.89. Furthermore, the minimum value of each corrected item-to-total correlation is above 0.5 (minimum = 0.517), suggesting that the instrument has good reliability, as shown in Table 1.

The content validity of a questionnaire refers to the representativeness of the item content domain. It is the manner in which the questionnaire and its items are built that ensures the reasonableness of the claims of content validity [24]. Churchill [7] suggested that "specifying the domain of the construct, generating items that exhaust the domain, and subsequently purifying the resulting scale should produce a measure which is content or face valid and reliable". Therefore, the rigorous procedures used to conceptualize the USQAS constructs based on previous research to form the initial items, the personal interviews with several experts, and the iterative procedures of the scale purification suggest that the USQAS instrument has strong content validity.

Criterion-related validity is defined as the effectiveness of a measure in predicting behavior in specific situations [24]. It is assessed by comparing the correlation coefficient test scores with the external criterion or overall satisfaction [13, 20]. In this study, we obtained the criterion-related validity by calculating 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 show that the 18-item USQAS instrument has a criterion-related validity of 0.62 and a significance level of 0.000, suggesting acceptable criterion-related validity.

The construct validity of a measure can be demonstrated by validating the theory behind the instrument [20, 31]. Construct validity is the extent to which an operational measure truly reflects the concept being investigated or the extent to which operational variables used to observe

covariation in and between constructs can be interpreted in terms of the model's theoretical constructs [30]. Researchers have used various validation strategies to establish construct validity, including item-to-total correlations [13, 20, 24], factor analysis, and assessment of convergent and discriminant validity [7, 13, 24, 35]. Convergent and discriminant validation demonstrates construct validity by showing that an instrument not only correlates with other variables with which it should correlate, but also that it does not correlate with variables from which it should differ[24].

We examined construct validity in terms of convergent and discriminant validity by using a correlation matrix approach [13, 31]. 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. Results show the correlation matrix of the measures. The smallest within-factor correlations are: ease of use = 0.72; usefulness = 0.57; service quality = 0.70; and content quality = 0.57. These correlations are significantly higher than 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 theoretical factor [13, 31, 40]. For example, the lowest within-factor correlation for usefulness is 0.57, however, one of the correlations of content with items of other factors is larger than 0.57. i.e., the number of violations is 1. For discriminant validity, Campbell and Fiske [5] suggested that the count should be less than 50% of the potential comparisons. The results show 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 this study.

## 6. Conclusion and Future Research

We have rigorously tested the USQAS instrument and found that it provides a high degree of confidence in the reliability and validity of the scales. A comprehensive model for measuring USQAS is presented in Figure 1. In this study, we developed 4-factor, 18-item instrument for measuring USQAS. The four primary dimensions of USQAS are ease of use, usefulness, service quality, and content quality.

To enhance user satisfaction and the success of question answering systems, we have developed an integrated theoretical evaluation model for such systems, based on a review and synthesis of existing IS user satisfaction and technology acceptance models. We believe the proposed evaluation model provides a framework for the design of question answering systems from the user's perspective and
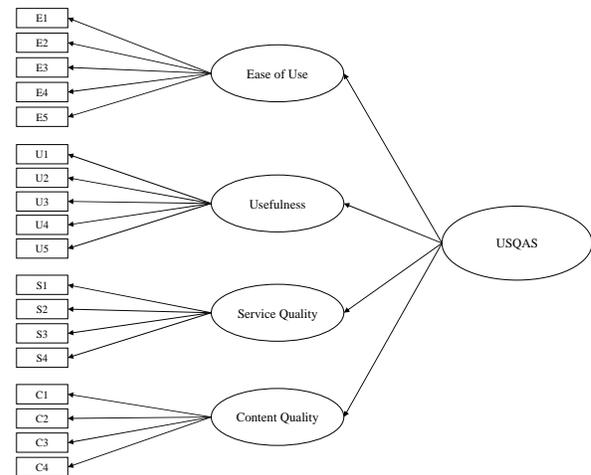


Figure 1. Proposed evaluation model for measuring user satisfaction with question answering systems (USQAS)

that it could help increase user acceptance of QAS.

Our study makes two contributions to the research and practitioner communities. First, it provides a framework for describing the primary dimensions of user satisfaction with QAS. Second, 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.

In the future, we will use the model to conduct an empirical study of question answering systems and analyze the theoretical and managerial implications of this study's results.

## 7. Acknowledgment

## 8. References

[1] S. F. Abdinnour-Helm, B. S. Chaparro, and S. M. Farmer, "Using the end-user computing satisfaction (EUCS) instrument to measure satisfaction with a Web site," *Decision Sciences,* vol. 36, pp. 341-364, May 2005.

[2] J. Allan, B. Carterette, and J. Lewis, "When Will Information Retrieval Be "Good Enough"? - User Effectiveness As a Function of Retrieval Accuracy," in *Proceedings of ACM SIGIR 2005*, 2005.

[3] J. E. Bailey and S. W. Pearson, "Development of a Tool for Measuring and Analyzing Computer User Satisfaction," *Management Science,* vol. 29, pp. 530-545, 1983.

[4] J. J. Baroudi and W. J. Orlikowski, "A Short-From Measure of User Information Satisfaction: A Psychometric Evaluation and Notes on Use," *Journal of Management Information Systems,* vol. 4, pp. 44-59, Spring 1988 1988.

[5] D. T. Campbell and D. W. Fiske, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin,* vol. 56, pp. 81-105, 1959.

[6] W. Chen, Q. T. Zeng, W. Y. Liu, and T. Y. Hao, "A user reputation model for a user-interactive question answering system," *Concurrency and Computation-Practice & Experience,* vol. 19, pp. 2091-2103, Oct 2007.

[7] G. A. Churchill, "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research,* vol. 16, pp. 64-73, 1979.

[8] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika,* vol. 16, pp. 297-334, 1951.

[9] R. D. Davis and R. P. Bagozzi, "User acceptance of computer technology: a comparison of two theoretical models," *Management Science,* vol. 35, pp. 982-1003, 1989.

[10] W. H. DeLone and E. R. McLean, "The DeLone and McLean Model of Information Systems Success: A Ten-Year Update," *Journal of Management Information Systems,* vol. 19, pp. 9-30, Spring 2003.

[11] S. Devaraj, M. Fan, and R. Kohli, "Antecedents of B2C channel satisfaction and preference: Validating e-commerce metrics," *Information Systems Research,* vol. 13, pp. 316-333, Sep 2002.

[12] A. R. Diekema, O. Yilmazel, and E. D. Liddy, "Evaluation of Restricted Domain Question-Answering Systems," in *Proceedings of EACL Workshop on Question Answering in Restricted Domains*, Barcelona, Spain, 2004.

[13] W. J. Doll and G. Torkzadeh, "The Measurement of End-User Computing Satisfaction," *MIS Quarterly,* vol. 12, pp. 259-274, Jun 1988.

[14] W. J. Doll and G. Torkzadeh, "The Measurement of End-User Computing Satisfaction - Theoretical and Methodological Issues," *MIS Quarterly,* vol. 15, pp. 5-10, Mar 1991.

[15] W. J. Doll, X. D. Deng, T. S. Raghunathan, G. Torkzadeh, and 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,* vol. 21, pp. 227-262, Sum 2004.

[16] M. Fishbein and I. Ajzen, *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. MA: Addison-Wesley, 1975.

[17] A. W. Gatian, "Is User Satisfaction a Valid Measure of System Effectiveness," *Information & Management,* vol. 26, pp. 119-131, Mar 1994.

[18] M. Gelderman, "The relation between user satisfaction, usage of information systems and performance," *Information & Management,* vol. 34, pp. 11-18, Aug 5 1998.

[19] J. E. Hair and R. L. Anderson, *Multivariate data analysis*. Upper Saddle River, NJ: Prentice-Hall, 1998.

[20] B. Ives, M. H. Olson, and J. J. Baroudi, "The measurement of user information satisfaction," *Communications of the ACM,* vol. 26, pp. 785-793, 1983.

[21] T. Kato, J. Fukumoto, F. Masui, and N. Kando, "Are open-domain question answering technologies useful for information access dialogues? - An empirical study and a proposal of a novel challenge," *ACM Transactions on Asian Language Information Processing,* vol. 4, pp. 243-262, 2005.

[22] D. Kelly, P. B. Kantor, E. L. Morse, J. Scholtz, and Y. Sun, "User-Centered Evaluation of Interactive Question Answering Systems," in *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL*, 2006, pp. 49-56.

[23] D. Kelly, N. Wacholder, R. Rittman, Y. Sun, P. Kantor, S. Small, and T. Strzalkowski, "Using interview data to identify evaluation criteria for interactive, analytical question-answering systems," *Journal of the American Society for Information Science and Technology,* vol. 58, pp. 1032-1043, May 2007.

[24] K. Klenke, "Construct Measurement in Management-Information-Systems - a Review and Critique of User Satisfaction and User Involvement Instruments," *INFOR,* vol. 30, pp. 325-348, Nov 1992.

[25] T. Kokubu, T. Sakai, Y. Saito, H. Tsutsui, T. Manabe, M. Koyama, and H. Fujii, "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.

[26] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. R. Karger, "What Makes a Good Answer? The Role of Context in Question Answering," in *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003)*, Zurich, Switzerland, 2003.

[27] J. Lin and D. Demner-Fushman, "Methods for automatically evaluating answers to complex questions," *Information Retrieval,* vol. 9, pp. 565-587, 2006.

[28] J. Lin, "User simulations for evaluating answers to question series," *Information Processing & Management,* vol. 43, pp. 717-729, May 2007.

[29] F. Lopez-Ostenero, V. Peinado, J. Gonzalo, and F. Verdejo, "Interactive question answering: Is Cross-Language harder than monolingual searching?," *Information Processing & Management,* vol. 44, pp. 66-81, Jan 2008.

[30] R. G. Netemeyer, W. O. Bearden, and S. Sharma, *Scaling Procedures: Issues and Applications*. Thousand Oaks, CA: Sage Publications, 2003.

[31] C. S. Ong and J. Y. Lai, "Measuring user satisfaction with knowledge management systems: scale development, purification, and initial test," *Computers in Human Behavior,* vol. 23, pp. 1329-1346, May 2007.

[32] R. P. Schumaker and H. Chen, "Leveraging Question Answer technology to address terrorism inquiry," *Decision Support Systems,* vol. 43, pp. 1419-1430, Aug 2007.

[33] R. P. Schumaker, M. Ginsburg, H. C. Chen, and Y. Liu, "An evaluation of the chat and knowledge delivery components of a low-level dialog system: The AZ-ALICE experiment," *Decision Support Systems,* vol. 42, pp. 2236-2246, Jan 2007.

[34] R. P. Schumaker, Y. Liu, M. Ginsburg, and H. C. Chen, "Evaluating the efficacy of - A terrorism question/answer system," *Communications of the Acm,* vol. 50, pp. 74-80, Jul 2007.

[35] D. W. Straub, "Validating Instruments in MIS Research," *MIS Quarterly,* vol. 13, pp. 147-169, Jun 1989.

[36] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seatttle, WA, 2006, pp. 11-18.

[37] V. Venkatesh and F. D. Davis, "A model of the antecedents of perceived ease of use: Development and test," *Decision Sciences,* vol. 27, pp. 451-481, Sum 1996.

[38] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward A Unified View," *MIS Quarterly,* vol. 27, p. 425, Sep 2003.

[39] N. Wacholder, D. Kelly, P. Kantor, R. Rittman, Y. Sun, B. Bai, S. Small, B. Yamrom, and T. Strzalkowski, "A model for quantitative evaluation of an end-to-end question-answering system," *Journal of the American Society for Information Science and Technology,* vol. 58, pp. 1082-1099, Jun 2007.

[40] Y. S. Wang, H. Y. Wang, and D. Y. Shee, "Measuring e-learning systems success in an organizational context: Scale development and validation," *Computers in Human Behavior,* vol. 23, pp. 1792-1808, Jul 2007.

[41] B. H. Wixom and P. A. Todd, "A theoretical integration of user satisfaction and technology acceptance," *Information Systems Research,* vol. 16, pp. 85-102, Mar 2005.

[42] J. H. Wu, S. C. Wang, and L. M. Lin, "Mobile computing acceptance factors in the healthcare industry: A structural equation model," *International Journal of Medical Informatics,* vol. 76, pp. 66-77, Jan 2007.

[43] H. Yu, M. Lee, D. Kaufman, J. Ely, J. A. Osheroff, G. Hripcsak, and J. Cimino, "Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians," *Journal of Biomedical Informatics,* vol. 40, pp. 236-251, Jun 2007.