Internet Data Quality: Perceptions of Graduate and Undergraduate Business Students

Barbara D. Klein, University of Michigan at Dearborn

This study examines user perceptions of the quality of information found on the Internet using an instrument that builds on prior research identifying dimensions of data quality. Surveys of graduate and undergraduate students taking MIS courses were conducted. Differences were found in the graduate and undergraduate student ratings for four dimensions of Internet data quality. The results suggest that at least some users are aware of the relative strengths and weaknesses of information published on the Internet and that educational experiences have some effect on perceptions of the quality of information published on the Internet.

INTRODUCTION

Data quality problems may occur because of the ease with which information can be published on the Internet. For example, peer review and editorial processes are sometimes missing when information is made available through the Internet. Although there is anecdotal evidence of problems with the quality of information available through Internet sources (Calishain, 1997), little research measuring users' perceptions of the quality of this information has been conducted. A study of student perceptions of the quality of information retrieved using the Internet was conducted to address this research gap. A survey study of graduate student and undergraduate student perceptions of the quality of information retrieved from the Internet is reported in this paper. The study applies the considerable research that has been done on the dimensions of data quality. The remaining sections of the paper discuss (1) data quality and the Internet, (2) research on the dimensions of data quality, (3) a survey instrument measuring data quality from the perspective of users of data, (4) the research propositions, (5) the research methodology, and (6) the empirical results.

DATA QUALITY AND THE INTERNET

Problems may occur with information published on the Internet because review processes are often absent (Hawkins, 1999; Pack, 1999). Fuld (1998) warns executives of the dangers of old data and irrelevant information and notes that poor information quality on the Internet can damage business performance. Keltner (1998) similarly cautions against problems with data quality in the context of using the Internet as a medium for publishing consumer-oriented health information.

Several authors have recognized these data quality problems and proposed checklists and frameworks providing prescriptive advice for evaluating the quality of information published on the Internet. Hawkins (1999) discusses fourteen criteria that are included in these

prescriptive checklists. The criteria include (1) currency/updating, (2) purpose/author/bias, (3) author/source, (4) scope, (5) accuracy/relevance, (6) design/format, (7) authority, (8) uniqueness/stability, (9) structure/indexing, (10) review/ratings, (11) writing quality, (12) data quality, (13) selection criteria, and (14) links to/from other sources. Alexander and Tate (1999) suggest five criteria for evaluating Internet-based information: authority, accuracy, currency, objectivity, and coverage. Pack (1999) reviews other resources available on the Internet that offer guidance for evaluating information on the Internet.

Despite the prescriptive frameworks developed in recent years for evaluating the quality of Internet-based information, little is known about users' actual perceptions of the quality of Internet-based information. However, in one study the web was found to be perceived as less authoritative and credible than other types of information systems (Rieh & Belkin, 1998). Another study compared user perceptions of the quality of information retrieved from Internet and traditional text sources (i.e., books, journals, magazines, and newspapers). In this study, traditional text sources were rated as more accurate and objective than Internet sources. Users also found the reputation of traditional text sources to be better than that of Internet sources and the formatting of traditional text sources to be more consistent than that of Internet sources. In contrast, Internet sources were rated higher in terms of their timeliness and amount of available data (Klein, 2001).

The objective of the study reported in this paper is to further improve our understanding of users' evaluations of Internet information quality by comparing perceptions of graduate and undergraduate students. The study is built on prior research aimed at understanding the dimensions of data quality.

DIMENSIONS OF DATA QUALITY

Data quality is typically conceptualized as a multi-dimensional concept. For example, in an early discussion of the quality of information systems, Davis and Olson (1985) identify three aspects of data quality: accuracy, precision, and completeness

Huh, Keller, Redman, and Watkins (1990) define four dimensions of data quality: accuracy, completeness, consistency, and currency. They define accuracy as agreement with either an attribute about a real world entity, a value stored in another database, or the result of an arithmetic computation. Completeness is defined with respect to some specific application and refers to whether all of the data relevant to that application are present. Consistency refers to an absence of conflict between two datasets. Currency refers to whether the data are up-to-date.

Fox, Levitin, and Redman (1993) also identify these four dimensions of data quality (accuracy, completeness, consistency, and currency). Accuracy refers to whether a data value matches some value considered to be correct. Completeness means that a collection of data contains values for all fields that should have values and that no records are missing. Consistency refers to whether data values conform to constraints that have been specified for that data. Currency refers to whether a data value is up-to-date

In contrast to these conceptual frameworks, Zmud (1978) and Madnick and Wang (1992) present definitions of data quality derived from empirical observation. Zmud (1978) used factor analysis to examine the dimensionality of the construct of information. Four dimensions were derived: quality of information, relevancy of information, quality of format, and quality of meaning. Madnick and Wang (1992) use observations of defective data in organizational databases to derive four components of data quality: completeness, accuracy, appropriateness, and consistency.

Wand and Wang (1996) argue for a definition of data quality that is task-independent and identify four dimensions of intrinsic data quality: completeness, lack of ambiguity, meaningfulness, and correctness. These dimensions are said to be applicable across different applications applied to different tasks.

A USER-DRIVEN VIEW OF DATA QUALITY

Wang and Strong (1996) develop a framework of dimensions of data quality from the perspective of users of the data. Surveys of users were conducted to generate a comprehensive list of data attributes. In one survey, users were asked to list attributes of data quality, and a total of 118 attributes were generated. In a second survey, users rated the importance of these 118 data attributes. An exploratory factor analysis of their responses was performed, and twenty dimensions of data quality were extracted. The twenty dimensions exhibited stability across a series of factor analyses in which the number of required factors was changed. The dimensions also had face validity and demonstrated acceptable reliability with Cronbach's alpha between .69 and .98 (Wang & Strong, 1996).

Because the researchers felt that an instrument based on twenty dimensions could be unwieldy in practice, they conducted an additional study in which a different group of subjects sorted these twenty dimensions into four conceptually-derived categories (accuracy, relevancy, representation, and accessibility). Fifteen dimensions (encompassing 50 data attributes) emerged from the sorting process. The fifteen dimensions are believability, accuracy, objectivity, reputation, value-added, relevancy, timeliness, completeness, appropriate amount of data, interpretability, ease of understanding, representational consistency, concise representation, accessibility, and access security (Wang & Strong, 1996). The fifteen dimensions and the data attributes associated with each dimension are listed in Appendix A.

The resulting framework is proposed as a tool for measuring data quality (Wang & Strong, 1996). Strong, Lee, and Wang (1997) use the framework to discuss data quality problems in three organizations.

RESEARCH QUESTIONS AND PROPOSITIONS

Three research questions are examined in this study.

1. What dimensions of Internet data quality are rated highest and lowest by graduate business students?

2. What dimensions of Internet data quality are rated highest and lowest by undergraduate business students?

3. What are the differences between graduate business student and undergraduate business student perceptions of Internet data quality?

The Wang and Strong (1996) framework is applied in this study as a tool for measuring data quality. This framework was selected as the foundation for this study for a number of reasons. First, it provides a very comprehensive look at data quality from the perspective of users of data. Second, it was developed using a well-established methodology for determining aspects of the quality of various products. Third, it provides a validated instrument for assessing data quality. And finally, the framework has proven valuable in prior applications aimed at identifying and solving data quality problems in Fortune 100 companies and in the United States military (Wang & Strong, 1996)

Using this framework, fifteen research propositions (each corresponding to one of the dimensions of data quality) about differences in perceptions of the quality of Internet-based information are examined. The research proposition for the believability dimension is stated below.

Proposition: *There is no difference in graduate business student and undergraduate business student perceptions of the believability of information available from the Internet.*

The propositions for the other fourteen dimensions of data quality have the same form as this proposition.

RESEARCH METHODOLOGY

This study examines evaluations of the quality of information provided through the Internet from the perspective of the users of this data. Surveys were administered to two groups of students. First, a survey was administered to 55 graduate students taking an MBA course following the completion of a course project requiring the use of the Internet. Sixty-two students were eligible to complete the graduate student survey, giving a response rate of 89 percent. Second a survey was administered to 57 undergraduate students taking an MIS course following the completion of a course project requiring the use of the Internet. Sixty-five students were eligible to complete the undergraduate student survey, giving a response rate of 88 percent.

For both groups of students, questions about the extent to which the fifty data attributes identified by Wang and Strong (1996) describe data from Internet sources used for the course project were asked. The survey included fifty items. All questions were asked in the context of the course project that respondents conducted prior to completing the survey. A sample question is shown below for one of the data attributes (accuracy).

Data used for the course project from Internet sources were accurate.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

EMPIRICAL RESULTS

Table 1 shows the mean scores for each of the fifteen data quality dimensions. Responses for each survey item could range from one to seven. Higher scores reflect more favorable perceptions of data quality. For dimensions with more than one data attribute, scores are averaged across the data attributes. Results are presented separately for the graduate and undergraduate students. Statistically significant ($p<.05$) differences between the graduate and undergraduate student perceptions are indicated in the last column of the table.

TABLE 1
Mean Scores for Data Quality Dimensions
(See Appendix B for mean and standard deviations)

Dimension of Data Quality	Graduate Student Perceptions	Undergraduate Student Perceptions	Significant Difference (at $p<.05$)
Believability	5.60	5.53	No
Accuracy	4.17	4.52	Yes
Objectivity	4.03	4.21	No
Completeness	4.70	5.07	No
Reputation	4.61	4.70	No
Value-Added	4.80	5.04	No
Relevancy	5.38	5.18	No
Timeliness	5.51	5.07	No
Appropriate Amount	5.35	5.14	No
Interpretability	5.24	5.05	No
Ease of Understanding	5.04	5.19	No
Representational Consistency	3.87	4.50	Yes
Concise Representation	4.39	4.84	Yes
Accessibility	5.37	5.36	No
Access Security	2.61	3.49	Yes

Discussion of the results will begin with an analysis of the highest and lowest rated data quality dimensions for the graduate and undergraduate students. Next, a comparison of perceptions of the graduate and undergraduate students is presented.

Graduate Student Perceptions of Internet Data Quality. For the graduate students, mean scores across the fifteen data quality dimensions ranged from 5.60 to 2.61. Table 2 shows the five highest rated and the five lowest rated dimensions for the graduate students.

Timeliness and accessibility are among the highest rated data quality dimensions for the graduate students. Presumably, this reflects the speed with which information can be published on the Internet and the ease with which information can be accessed on the Internet. Other highly rated dimensions for Internet sources include believability, relevancy, and appropriate amount.

TABLE 2
Highest and Lowest Rated Data Quality Dimensions for Graduate Students

Highest Rated Data Quality Dimensions		Lowest Rated Data Quality Dimensions	
Believability	5.60	Access Security	2.61
Timeliness	5.51	Representational Consistency	3.87
Relevancy	5.38	Objectivity	4.03
Accessibility	5.37	Accuracy	4.17
Appropriate Amount	5.35	Concise Representation	4.39

Accuracy and objectivity received low ratings. Users of Internet-based information seem to be aware that errors may be present in this information. Two other dimensions with low ratings for Internet-based information are concise representation and representational consistency. These ratings reflect problems with the formatting of Internet-based information. Finally, access security was the lowest rated dimension for Internet-based information. Presumably, this reflects users' perceptions of the open access that characterizes the Internet.

Undergraduate Student Perceptions of Internet Data Quality. For the undergraduate students, mean scores across the fifteen data quality dimensions ranged from 5.53 to 3.49. Table 3 shows the five highest rated and the five lowest rated dimensions for the undergraduate students.

Believability, accessibility, ease of understanding, relevancy, and appropriate amount are the highest rated data quality dimensions for the undergraduate students. Like the graduate students, these students are aware of the ease with which information can be accessed on the Internet. The undergraduate students gave their lowest ratings to access security, objectivity, representational consistency, accuracy, and reputation. These users are also aware that errors may be present in Internet-based information.

Comparison of Graduate and Undergraduate Student Perceptions. Table 1 indicates that statistically significant differences ($p<.05$) were found between the ratings of the graduate and undergraduate students for four of the data quality dimensions: accuracy, representational

TABLE 3
Highest and Lowest Rated Data Quality Dimensions for Undergraduate Students

Highest Rated Data Quality Dimensions		Lowest Rated Data Quality Dimensions	
Believability	5.53	Access Security	3.49
Accessibility	5.36	Objectivity	4.21
Ease of Understanding	5.19	Representational Consistency	4.50
Accessibility	5.18	Accuracy	4.52
Appropriate Amount	5.14	Reputation	4.70

consistency, concise representation, and access security. The undergraduate students gave higher ratings on average for all four of the dimensions for which significant differences were found. Interestingly, no statistically significant differences were found for the other eleven data quality dimensions suggesting a large degree of agreement among users with different educational backgrounds.

CONCLUSION

During the past decade, we have witnessed an explosion in the amount of information available through the Internet. This study improves our understanding of how different groups of users perceive the quality of this information.

The evidence collected in this study suggests that users recognize that the Internet gives them easy access to appropriate amounts of believable, relevant information. On the other hand, users recognize that problems with accuracy, objectivity, and access security occur with Internet-based information. Differences in graduate and undergraduate student perceptions were noted for four data quality dimensions suggesting that education has some effect on users' perceptions of Internet-based information. Compared to the undergraduate students, the more educated group of students found Internet-based material to be less accurate, less secure, and more poorly presented and formatted. These lower ratings are arguably reflective of objective characteristics of information published on the Internet that has not been subjected to review and editorial processes.

The study has several limitations. First, the sample size was relatively small. Second, all of the surveyed users were students taking MIS courses. Compared to other students, these students may have a more sophisticated understanding of the strengths and weaknesses of Internet-based information. Third, respondents were asked questions about the quality of Internet-based information used for the class project in general rather than being asked questions about specific Internet sites.

Nevertheless, the findings of this study suggest that additional research in this area would be beneficial. Future research should focus on surveying a larger, less narrowly focused user population and on surveying users about their perceptions of the data quality of specific web sites.

REFERENCES

Alexander, J. E., & Tate, M. A. (1999). *Web wisdom: How to evaluate and create information quality on the web.* Mahwah, NJ: Lawrence Erlbaum Associates.

Calishain, T. (1997). *Official Netscape guide to Internet research.* Research Triangle Park, NC: Ventana Communications Group, Inc.

Davis, G. B., & Olson, M. H. (1985). *Management information systems: Conceptual foundations, structure, and development.* New York: McGraw-Hill Book Company.

Fox, C., Levitin, A., & Redman, T. (1993). The notion of data and its quality dimensions. *Information Processing & Management, 30,* 9-19.

Fuld, L. M. (1998). The danger of data slam. *CIO Enterprise Magazine (Sept.15),* 28-33. Online <http://www.cio.com/archive/enterprise/091598_ic.html>

Hawkins, D. T. (1999). What is credible information? *Online, 23*(5), 86-89.

Huh, Y. U., Keller, F. R., Redman, T. C., & Watkins, A. R. (1990). Data quality. *Information and Software Technology, 32*, 559-565.

Keltner, K. B. (1998). Networked health information: Assuring quality control on the Internet. *Federal Communications Law Journal, 50*(2), 417-439.

Klein, B. D. (2001). User perceptions of data quality: Internet and traditional text sources. *Journal of Computer Information Systems, 41*(4), 9-15.

Madnick, S. E., & Wang, R. Y. (1992). *Introduction to the TDQM research program*. Total Data Quality Management Research Program Working Paper #92-01.

Pack, T. (1999). Can you trust Internet information? *Link-up, 16*(6), 24.

Rieh, S. Y., & Belkin, N. J. (1998). Understanding judgment of information quality and cognitive authority in the WWW. *Journal of the American Society for Information Science, 35*, 279-289.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103-110.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86-95.

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

Zmud, R. W. (1978). An empirical investigation of the dimensionality of the concept of information. *Decision Sciences, 9*, 187-195.