AMCIS 2005 Proceedings

Americas Conference on Information Systems (AMCIS)

2005

An Empirical Investigation of the Impact of Data Quality and its Antecedents on Data Warehousing

Nicole Lang Beebe
University of Texas at San Antonio, nicole.beebe@utsa.edu

Diane Walz
University of Texas at San Antonio, Diane.Walz@utsa.edu

An Empirical Investigation of the Impact of Data Quality and its Antecedents on Data Warehousing Success

Nicole Lang Beebe
University of Texas at San Antonio
nicole.beebe@utsa.edu

Diane Walz
University of Texas at San Antonio
diane.walz@utsa.edu

ABSTRACT

Data warehousing is a topic of great interest in the business community, due to increasing business intelligence demands, coupled with increased data availability and processing capability. Despite large financial backing of data warehousing implementations, many fail. Little research has been conducted pertaining to data warehousing success. Traditional system success models (DeLone and McLean, 1992; Seddon, 1997) may be extensible to data warehousing, provided both infrastructure and business application aspects of the implementation are carefully considered, and provided increased attention is paid to the antecedents of data and system quality. Wixom and Watson (2001) conducted an empirical study examining the antecedents to data and system quality to data warehousing success, but found no statistically significant support for the data quality antecedents proposed. This paper reviews system success and data quality literature and proposes a new model for data warehousing success. The new model extends traditional system success models to data warehousing, but proposes a new set of data quality antecedents, which can be empirically examined.

Keywords

Data warehouse, data quality, information systems success.

INTRODUCTION & BACKGROUND

Data warehousing has emerged as an important topic in information systems (IS) since the mid- to late 1990s (Wixom and Watson, 2001), due to its increasing importance to corporations. Its increased importance is a function of growing business intelligence demands and significant technological advances that have increased overall data availability, storage capacity, and processing capability (Negash, 2004). Data warehousing and data mining are considered essential components of proactive business intelligence (Langseth and Vivatrat, 2002). Benefits of data warehousing include time savings, improved information, improved decision making, data mart consolidation, personnel savings, business process improvements, and support for strategic business objectives (Watson, Abraham, Chen, Preston and Thomas, 2004; Watson, Annino, Wixom, Avery and Rutherford, 2001).

Despite significant data warehousing demand (Watson et al., 2001) and technological advances, it is estimated that one-half to two-thirds of corporate data warehousing efforts fail (Wixom and Watson, 2001). In a survey of 106 organizations that implemented data warehouses, only 65.5% reported the data warehouse met its objectives (Watson et al., 2004). This is problematic given the cost for such investments, considering initial set-up costs and operating costs, averages in excess of one million dollars (Han and Kamber, 2001; Watson et al., 2004; Watson et al., 2001). Because of the obvious high risk associated with corporate data warehousing initiatives, research into the data warehousing success is particularly relevant.

Data warehousing arguably possesses both IS application and information technology (IT) infrastructure characteristics (Seddon, Staples, Patnayakuni and Bowtell, 1999; Wixom and Watson, 2001). Accordingly, traditional system success models (DeLone and McLean, 1992; Seddon, 1997) and findings may not extend directly to data warehousing. Few studies have examined IT infrastructure success (Duncan, 1995) or data warehouse success (Wixom and Watson, 2001). Wixom and Watson (2001) proposed a model for data warehouse success (see Figure 1), but found only partial empirical support for the model. They validated that data quality and system quality explain a combined thirty-seven percent (36.9%) of the variance in perceived net benefits, however their proposed data and system quality antecedents (see Figure 1) explained only 12.8% of system quality variance and *none of the variance in data quality*. Accordingly, Wixom and Watson (2001) acknowledged factors other than those proposed must inform data quality in a data warehouse implementation and recommended further research be conducted in that area.

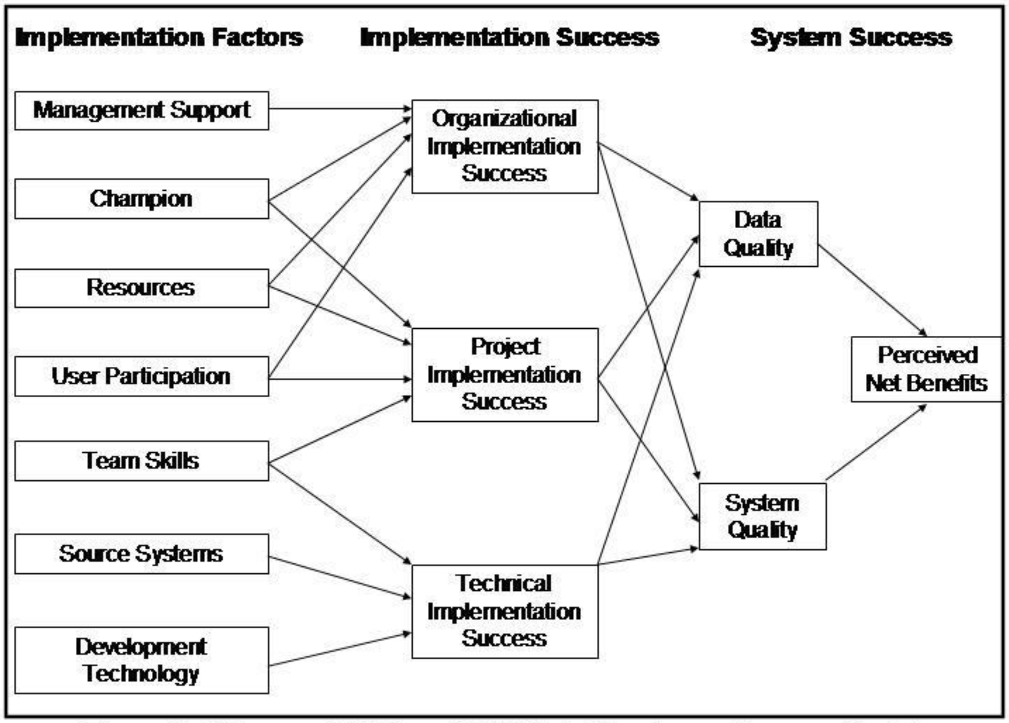


Figure 1. Wixom and Watson (2001) Data Warehouse Success Model

The purpose of this study is to better understand what makes data warehouses successful. To do so, we extend traditional system success models (DeLone and McLean, 1992; Seddon, 1997) to data warehousing, with specific emphasis on better understanding the antecedents to data quality in data warehousing implementations. In this effort, data quality is measured as a multi-dimensional construct and examined from both organizational and system perspectives.

PROPOSED RESEARCH MODEL

The proposed research model carries forward Seddon's (1997) system success model, wherein system quality and data quality are antecedents to perceived net benefits. To better understand data quality in data warehouses, however, Wixom and Watson's (2001) one-dimensional data quality measure is expanded to a three-dimensional measure, and different antecedents are proposed pursuant to extant data quality theory.

Information Systems Success

IS success has been operationalized in a variety of ways. In the DeLone and McLean (1992) model, IS success measures include use, user satisfaction, individual impact, and organizational impact. Seddon (1997) proposed a respecified model and argued that "IS use" is a proxy for "net benefits" that are derived from use. In both models, notions of system quality and information (data) quality are considered direct antecedents to IS system success. In line with previous research, including Wixom and Watson's (2001) research after which this research was modeled, the following is hypothesized:

H1a: Data quality is positively associated with perceived net benefits.

H1b: System quality is positively associated with perceived net benefits.

Data Quality and Antecedents

Data Quality Dimensionality

Data quality is particularly relevant to data warehousing success (Ballou and Tayi, 1999; Dijcks, 2004; Gonzales, 2004; Lee, Pipino, Strong and Wang, 2004; Redman, 1998). Wixom and Watson (2001) measured data quality in a uni-dimensional manner. Prior research, however, suggests data quality is not uni-dimensional. Capiello et al. (2003-4) suggest data quality is four-dimensional. Wand and Wang (1996) identify and empirically validate four dimensions of data quality, which map to three of the four dimensions proposed by Capiello et a. (2003-4). There appears to be consensus for the argument that data quality should be viewed and measured as a multi-dimensional construct.

To increase the explanatory power of the data quality construct in the overall data warehouse (DW) success model, the proposed model measures data quality along three dimensions: accuracy, relevancy, and accessibility. These dimensions leverage data quality taxonomy labels proffered by Wang and Strong (1996).

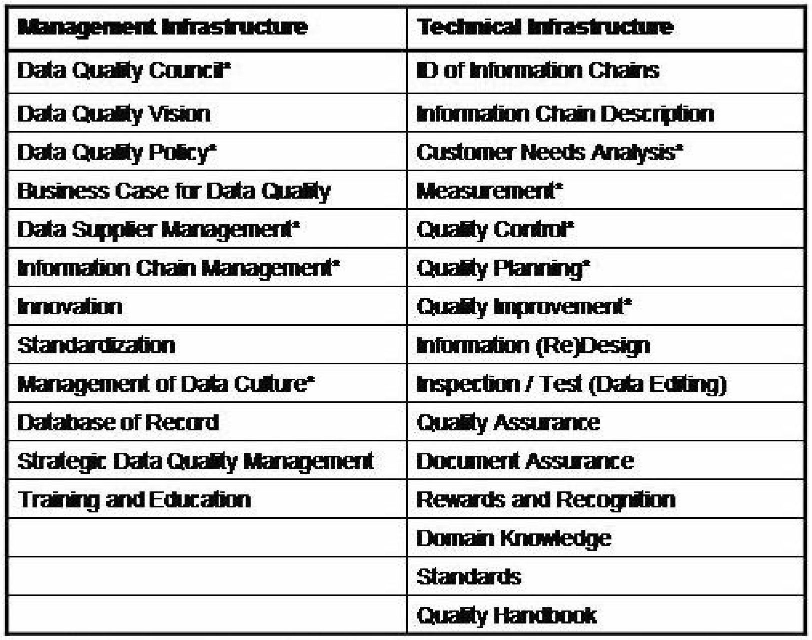
Data Quality Antecedents

Literature regarding the antecedents to data quality is scarce—considering both process models and variance models. Between the 1970's and early 1980's, data quality models focused on accounting reliability models that emphasized quality control of data entry and error checking in the form of ending financial balances (Ballou and Pazer, 1985). Ballou and Pazer (1985) then proposed a Data Flow/Data Processing Model that emphasized the ability of data processing to amplify, diminish, or not affect data errors.

In the mid-1980's, the *Total Quality Management (TQM)* movement began in America. In the context of TQM, data quality emerged as a deeper process issue, more so than just a data entry quality control issue regarding transaction processing systems (Redman, 1996, 1998, 2001). Redman (2001) defines data quality as follows: "Data are of high quality if they are fit for their intended uses in operation, decision making and planning" (p. 74).

Redman (2001) proposes a *data quality system (DQS)*, which is defined as "…the totality of an organization's efforts that bear on data quality" (p. 75). There are then two types of data quality systems employed in practice: *first generation data quality systems* and *second generation data quality systems.* First generation data quality systems address basic clean-up—finding and correcting data errors. Second generation data quality systems shift the focus to data error prevention.

In examining second generation data quality systems, Redman (2001) identifies twelve management infrastructure elements of second generation data quality systems, five of which are reportedly common to most successful data quality programs. He also identifies fifteen technical infrastructure elements of second generation data quality systems, five of which are reportedly common to most successful data quality programs. These data quality program elements are shown in Table 1 below.



Management Infrastructure	Technical Infrastructure
Data Quality Council*	ID of Information Chains
Data Quality Vision	Information Chain Description
Data Quality Policy*	Customer Needs Analysis*
Business Case for Data Quality	Measurement*
Data Supplier Management*	Quality Control*
Information Chain Management*	Quality Planning*
Innovation	Quality Improvement*
Standardization	Information (Re)Design
Management of Data Culture*	Inspection / Test (Data Editing)
Database of Record	Quality Assurance
Strategic Data Quality Management	Document Assurance
Training and Education	Rewards and Recognition
	Domain Knowledge
	Standards
	Quality Handbook

Table 1. Redman (2001) Data Quality Program Elements
***Elements common to most successful data quality programs**

Proposed antecedents to data warehouse data quality include both those elements recognized by Redman (2001) as common to most successful data quality programs, as well as four additional elements: Innovation, Standardization, Training & Education, and Inspection & Test (Data Editing), which appear particularly relevant in data warehousing.

Redman (2001) breaks these elements into two primary categories—*Management Infrastructure* and *Technical Infrastructure*. For clarity of understanding the constituent elements, we have renamed the primary categories: *Organizational Infrastructure* and *Data Quality System Implementation* respectively. These data quality element categories are defined as follows:

- Organizational Infrastructure: The degree to which senior management is committed to and involved with data quality, as well as the overall culture and characteristics of the organization which foster or hinder data quality.

- Data Quality System Implementation: The degree to which data quality is treated as a complete system with customer-driven requirements analysis; test, inspection, measurement, and evaluation processes; and data quality planning and improvement mechanisms.

Based on these definitions, four data quality antecedents are proposed that comprise these two data quality element categories. The four data quality antecedents are defined as follows:

- Management Commitment and Involvement: The degree to which and formalization with which senior management is committed to and involved with data quality. Such formalized commitment and involvement manifests itself in the form of policy, goal-setting, project initiation, accountability, data supplier management, improvement processes, and information chain management.

- Facilitating Organizational Culture and Characteristics: The widely held attitude that data and information are business assets and treated accordingly. Such treatment manifests itself in the form of innovation, aggressive data mining, appropriate use of standardization, minimal power struggles, and adequate training and education across the organization.

- Requirements Analysis and Technical Process Implementation: A data quality process that effectively identifies and handles customer needs and ensures quality control is maintained. Proper maintenance of quality control manifests itself in the form of data entry accuracy and empirical quality measurement and evaluation.

- Data Quality System Planning and Improvement: The proactive planning of data quality system implementation and incorporation of quality improvements therein.

The proposed data quality element structure is shown in Figure 2.

Based on these definitions, the following data quality antecedents are hypothesized:

H2a: “Management commitment & involvement” is positively associated with DW data quality.

H2b: “Facilitating organizational culture & characteristics” is positively associated with DW data quality.

H3a: “Requirements analysis & technical process implementation” is positively associated with DW data quality.

H3b: “Data quality system planning & improvement” is positively associated with DW data quality.

Finally, given the impact of first generation data quality on second generation data quality (Redman 2001), we propose consideration for the mediating effect of operational data quality between the proposed data quality antecedents and data warehouse data quality. As such, the following is hypothesized:

H4: Operational data quality within an organization is positively associated with DW data quality

H5a: “Management commitment & involvement” is positively associated with operational data quality.

H5b: “Facilitating organizational culture & characteristics” is positively associated with operational data quality.

H6a: “Requirements analysis & technical process implementation” is positively associated with operational data quality.

H6b: “Data quality system planning & improvement” is positively associated with operational data quality.

These hypotheses are reflected in the full research model shown in Figure 3.

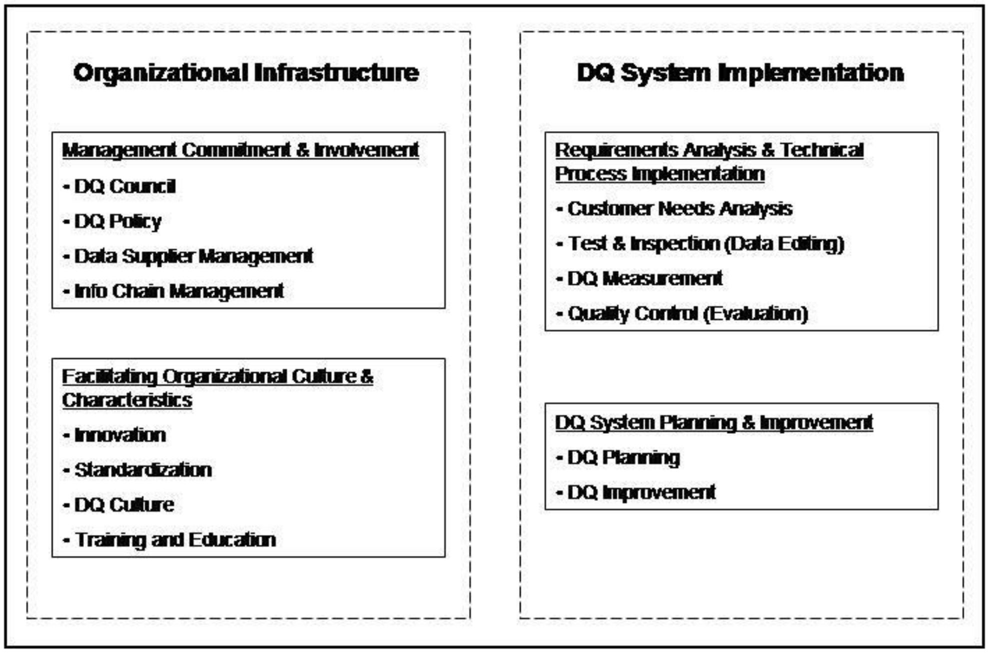

Organizational Infrastructure
Management Commitment & Involvement
- DQ Council
- DQ Policy
- Data Supplier Management
- Info Chain Management
Facilitating Organizational Culture & Characteristics
- Innovation
- Standardization
- DQ Culture
- Training and Education
DQ System Implementation
Requirements Analysis & Technical Process Implementation
- Customer Needs Analysis
- Test & Inspection (Data Editing)
- DQ Measurement
- Quality Control (Evaluation)
DQ System Planning & Improvement
- DQ Planning
- DQ Improvement

Figure 2. Proposed Data Quality Element Structure

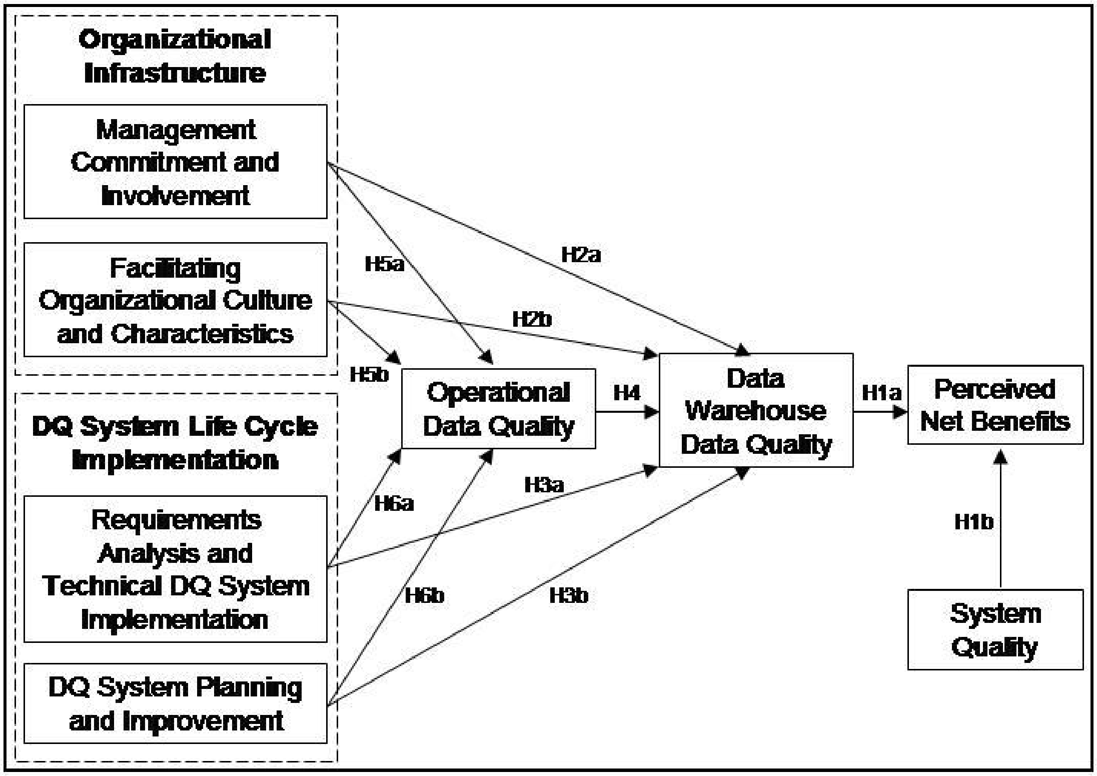

Organizational Infrastructure
Management Commitment and Involvement
Facilitating Organizational Culture and Characteristics
DQ System Life Cycle Implementation
Requirements Analysis and Technical DQ System Implementation
DQ System Planning and Improvement
Operational Data Quality
Data Warehouse Data Quality
Perceived Net Benefits
System Quality
H5a
H2a
H2b
H5b
H4
H1a
H1b
H6a
H3a
H6b
H3b

Figure 3. Proposed Research Model

PROPOSED RESEARCH METHOD

Systems success measures will be replicated from the Wixom and Watson (2001) study, which were derived from previously validated instruments. Data quality antecedent measures will be developed internally and subsequently tested for reliability and validity[1]. The final instrument containing system success and data quality measures will consist of a two-part, written, self-report survey using a nine-point Likert scale for all questions. All questions will be worded positively and the respondent will be asked to respond regarding their level of agreement/disagreement with system success measures and the extent to which stated data quality antecedents have been implemented.

The sample frame will consist of organizations (government agencies, commercial companies, consultants, and data warehousing vendors) internationally who have been known to have implemented operational data warehouses. The survey will be mailed and follow-up contact will be made as appropriate. The two survey parts will be handled separately within the organization. The system success instrument will be administered to knowledge workers—the users of the data warehouse. The data quality antecedents (elements) instrument will be administered to data warehouse managers.

The research model will be tested primarily using Partial Least Squares (PLS), a structural equation modeling technique. PLS will be used because the sample size will presumably be small, and because the model includes both formative and reflective measures. Reliability and validity estimations will also be calculated during the application of PLS, and subsequently reviewed. Finally, *t*-tests will be conducted to test for potential confounding variables, such as time since the data warehouse was implemented, size of organization, etc.

CONCLUSION

This is a work in progress. Research is currently on-going. Anticipated challenges of the proposed research method include access to the sample frame and successful development of a data quality antecedent measure with acceptable levels of construct validity. Extensive pre- and pilot-testing will be conducted to overcome the latter challenge. Success in overcoming the former challenge will be facilitated by participation in The Data Warehousing Institute's Conference(s).

REFERENCES

1. Ballou, D. P. and Pazer, H. L. (1985) Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems, *Management Science,* 2, 31, 150-164.
2. Ballou, D. P. and Tayi, G. K. (1999) Enhancing Data Quality in Data Warehouse Environments, *Communications of the ACM,* 1, 42, 73-79.
3. Cappiello, C., Francalanci, C. and Pernici, B. (2003-4) Time-Related Factors of Data Quality in Multichannel Information Systems, *Journal of Management Information Systems,* 3, 20, 71-91.
4. DeLone, W. H. and McLean, E. R. (1992) Information Systems Success: The Quest for the Dependent Variable, *Information Systems Research,* 1, 3, 60-96.
5. Dijcks, J.-P. (2004) Integrating Data Quality into Yor Data Warehouse Architecture, *Business Intelligence Journal,* 2, 9, 18.
6. Duncan, N. B. (1995) Capturing Flexibility of Information Technology Infrastructure: A Study of Resource Characteristics and Their Measure, *Journal of Management Information Systems,* 2, 12, 37-58.
7. Gonzales, M. L. (2004) The Data Quality Audit, *Intelligent Enterprise,* 11, 7, 18-21.
8. Han, J. and Kamber, M. (2001) Data Mining: Concepts and Techniques, Academic Press, San Diego.
9. Langseth, J. and Vivatrat, N. (2002) Why Proactive Business Intelligence is a Hallmark of the Real-Time Enterprise: Outward Bound, *Intelligent Enterprise,* 18, 5, 34-41.
10. Lee, Y. W., Pipino, L., Strong, D. M. and Wang, R. Y. (2004) Process-Embedded Data Integrity, *Journal of Database Management,* 1, 15, 87-104.

[1] Instrument development is on-going. Contact the authors for further information.

11. Negash, S. (2004) Business Intelligence, *Communications of the Association for Information Systems,* 13, 33.

12. Redman, T. C. (1996) Data Quality for the Information Ag, Artech House, Boston.

13. Redman, T. C. (1998) The Impact of Poor Data Quality on the Typical Enterprise, *Communications of the ACM,* 2, 41, 79-83.

14. Redman, T. C. (2001) Data Quality: The Field Guid, Digital Press, Boston.

15. Seddon, P. B. (1997) A Respecification and Extension of the DeLone and McLean Model of IS Success, *Information Systems Research,* 3, 8, 240-254.

16. Seddon, P. B., Staples, S., Patnayakuni, R. and Bowtell, M. (1999) Dimensions of Information Systems Success, *Communications of the Association for Information Systems,* 20, 2, 1-61.

17. Wand, Y. and Wang, R. Y. (1996) Achoring Data Quality Dimensions in Ontological Foundations, *Communications of the ACM,* 11, 39, 86-96.

18. Wang, R. Y. and Strong, D. M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems,* 4, 12, 5-34.

19. Watson, H. J., Abraham, D. L., Chen, D., Preston, D. and Thomas, D. (2004) Data Warehousing ROI: Justifying and Assessing a Data Warehouse, *Business Intelligence Journal*, 2, 9, 6-18.

20. Watson, H. J., Annino, D. A., Wixom, B., Avery, K. L. and Rutherford, M. (2001) Current Practices in Data Warehousing, *Information Systems Management*, 18, 5, 47-55.

21. Wixom, B. and Watson, H. J. (2001) An Empirical Investigation of the Factors Affecting Data Warehousing Success, *MIS Quarterly,* 1, 25, 17-41.