Dimensions of Data Quality:
Toward Quality Data by Design

Y. Richard Wang
Lisa M. Guarascio

August 1991
IFSRC Discussion Paper #CIS-91-06

Composite Information Systems Laboratory
E53-320, Sloan School of Management
Massachusetts Institute of Technology
30 Wadsworth Street
Cambridge, Mass. 02139
ATTN: Prof. Richard Wang
Tel. (617) 253-0442
Fax. (617) 734-2137
Bitnet Address: rwang@sloan.mit.edu

ACKNOWLEDGEMENTS Research conducted herein has been supported, in part, by MIT's International Financial Services Research Center. The authors wish to thank Professor France LeClerc for her advice on how to conduct the survey and analyze the survey results. Thanks are also due to Karen Lee for conducting field work and contributing ideas during this project, and Dae-Chul Sohn and Inseok Cha for their assistance.

Dimensions of Data Quality:
Toward Quality Data by Design

ABSTRACT As experience has shown, poor data quality can have serious social and economic consequences. Yet before one can address issues related to analyzing, managing and designing quality into data systems, one must first understand what data quality actually means. Furthermore, as is the case with manufacturing and sevice organizations, quality should be defined in relation to the consumer's needs and desires, not the producers. Thus, the focus of this paper is to identify the dimensions of data quality, as defined by actual data consumers, through well defined research methodologies instead of experience, anecdotes, and intuition. The end result of our research and analysis of data consumers yielded the following data quality dimensions.

(1) Believability
(2) Value Added
(3) Relevancy
(4) Accuracy
(5) Interpretability
(6) Ease of Understanding
(7) Accessibility
(8) Objectivity
(9) Timeliness
(10) Completeness
(11) Traceability
(12) Reputation
(13) Representational Consistency
(14) Cost Effectiveness
(15) Ease of Operation
(16) Variety of Data & Data Sources
(17) Concise
(18) Access Security
(19) Appropriate Amount of Data
(20) Flexibility

The most striking results of this analysis are that data quality means much more than just accuracy to data consumers, and that even accuracy is more complex than previously realized. Specifically, Believability, Value Added, and Relevancy were rated as more important to data consumers than accuracy, and data consumers valued the ability to trace data, the reputation of the data, and data source in order to assure themselves of the accuracy of the data. These dimension s can be applied to help analyze data quality and formulate quality data policy. More significantly, they can be used to establish a research foundation for the design of Quality Data Models and the development of Quality Data Base Management Systems.

Dimensions of Data Quality:
Toward Quality Data by Design

1. Introduction

Significant advances in the price, speed-performance, capacity, and capabilities of new database and telecommunication technologies have created a wide range of opportunities for corporations to align their information technology for competitive advantage in the marketplace. Across industries such as banking, insurance, retail, consumer marketing, and health care, the capabilities to access databases containing market, manufacturing, and financial information are becoming increasingly critical (Cash & Konsynski, 1985; Clemens, 1988; Goodhue, Quillard, & Rockart, 1988; Henderson, 1989; Ives & Learmonth, 1984; Keen, 1986; Madnick, Osborn, & Wang, 1990; Madnick & Wang, 1988; McFarlan, 1984).

It has been concluded, in a multi-year MIT research program, that corporations in the 1990s will integrate their business processes across traditional functional, product, and geographic lines. The integration of business processes, in turn, will accelerate demands for more effective application systems for product development, product delivery, and customer service and management (Morton, 1989; Rockart & Short, 1989). Increasingly, many important applications require access to corporate functional and product databases which have disparate levels of data qualities. Poor data quality, unfortunately, can have a substantial impact on corporate profits, as the literature reveals (Ballou & Tayi, 1989; Bodner, 1975; Hansen, 1983; Hansen & Wang, 1990; Laudon, 1986; Lindgren, 1991). We illustrate, in the following examples, the social and economic impact of data quality.

1.1 Data Quality: A Vital Social and Economic Issue

Credit reporting is one of the most striking examples of serious social consequences related to inaccurate data. The credit industry not only collects financial information on individuals, but also compiles employment records. The impact of an error on a credit report can be more devastating to an individual than merely the denial of credit. One congressional witness testified that "he lost his job when he was reported as having a criminal record...a record that really belonged to a man with a similiar name[1]." Another witness told how he had been plagued by bill collectors for over nine months: bill collectors who were trying to recover money owned by another man with the same name. In light of these testimonies, it is astonishing to learn from the New York Times and CBS evening news that Consumer's Union found that 48 percent of the credit reports that they surveyed contained errors, and 19 percent "had mistakes that could cause denial of credit, insurance or employment."[2]

When poor data quality results in poor customer service, there can be a direct negative impact on the corposrate bottom line. One of the largest providers of optical fiber in the world (Hansen & Wang, 1990) uses an automated computer system to mark fiber before shipment to customers because of the enormous variety of fiber produced. In early 1990, a *data accuracy* problem caused the system to mislabel a fiber shipment which subsequently was installed under a lake in the state of Washington. When the fiber malfunctioned, the company was forced to pay $500,000 for the removal of the cable, replacement of the experimental fibers, rebundling of the cable, and reinstallation of the cable. Although the company did everything it could to correct the problem, the damage to its reputation for customer service and quality was serious.

As another example, Boston City Hall discovered 6 million dollars worth of overcharges in their telephone bills over a period of years (Lindgren, 1991).

1.2 Research Focus and Significance

Before one can address issues involved in analyzing and managing data quality, one must first understand what data quality actually means. Just as it would be difficult to effectively manage a production line without understanding the attributes which define a quality product, it would also be difficult to analyze and manage data quality without understanding the attributes which define quality data.

The focus of this paper is to identify data quality dimensions through well-defined research methodologies instead of experience, anecdotes, and intuition. These dimensions, once defined, can be applied to help analyze data quality and formulate quality data policy.

[1] Source: Washington Post, June 9, 1991

[2] Source: New York Times, June 7, 1991

More significantly, it would establish a research foundation for the design of *Quality Data Models* and the development of *Quality Data Base Management Systems*. Modern database systems have been designed from the system perspective. Consequently, the integrity constraints and normalization theories (Maier, 1983), which are used to maintain the integrity and consistency of data stored in the database (Date, 1990), are necessary but not sufficient to attain the data quality demanded by non-system constituents.

1.3 Who Defines Data Quality?

The importance of looking at quality from the consumer's viewpoint has been stressed (Garvin, 1987):

> "To achieve quality gains, I believe, managers need a new way of thinking, a conceptual bridge to the consumer's vantage point. Obviously, market studies acquire a new importance in this context...One thing is certain: high quality means pleasing the consumer, not just protecting them from annoyances. (Garvin, 1987, p. 104)"

We chose to use Garvin's approach to defining data quality. That is, data quality is not defined by the producers or managers of data, such as Information Systems (IS) departments, but instead, is defined by the data consumer. Data quality, defined from this perspective, can be used by researchers and practitioners to direct their efforts toward quality data by design for data consumers instead of the IS professionals.

1.4 What is a Dimension?

For a manufacturing firm, the concept of quality encompasses much more than material defects. Garvin has developed a framework encompassing eight dimensions of quality: performance, features, reliability, conformance, durability, serviceability, aesthetics, and perceived quality (Garvin, 1988). Likewise, data quality encompasses much more than simply the *accuracy of data*. Thus, before we discuss specific data quality dimensions, we first must clarify what we consider to be the underpinnings of a data quality dimension.

We define a data quality dimension as a set of adjectives or characteristics which most data consumers react to in a fairly consistent way. That is, one thinks about the importance of all adjectives in the set in the same way, and this similarity holds across a majority of data consumers. For example, suppose that the adjectives *objective* and *unbiased* were grouped together from analysis and identified as a factor named *objectivity: objectivity* would be a dimension because most data consumers think of *objective* and *unbiased* to be part of the same dimension. In other words, if a person in strategic planning said *objectivity* was not important, then he/she would also consider *unbiased* not important. At the same time, a person in finance who considers *objectivity* crucial also would think that *unbiased* is crucial. Thus, a dimension is an underlying construct that data consumers use when evaluating data.

1.5 Paper Organization

Section 2 describes the research design, in particular the data analysis method, the generation of data quality attributes for identifying the dimensions of data quality, and the collection of data for uncovering the dimensions. Scetion 3 analyzes the data. We first present the descriptive statistics of the data. Next we present the factor analysis specifics and results. Based on the component loadings from the factor analysis, we define the dimensions uncovered and elaborate on each of these dimensions. Concluding remarks are made in Section 4.

2. Research Design

This section describes the method for data analysis, the generation of data quality attributes, and the collection of data for uncovering the dimensions of data quality.

2.1 Data Analysis Method

Upon preliminary analysis of the analytical tools and methods that are most commonly used to define consumer constructs and analyze data (Lehmann, 1989), we identified six methods: factor analysis, conjoint analysis, analysis of variance, cluster analysis, multidimensional scaling, and discriminant analysis. We chose to use factor analysis because one of the most frequent applications of factor analysis is to uncover an underlying data structure.

Factor analysis assumes that the surveyed variables are manifestations of a number of key, but unmeasured constructs. It then attempts to identify these underlying constructs by examining the relations among the responses to the surveyed variables. The rationale behind factor analysis is that the observed responses are actually produced by some unobserved factors. It is an ideal method for boiling down a large number of variables into a small number of factors. Since identifying key dimensions of data quality is our primary research goal, factor analysis is well-suited for our purposes.

Mathematically, factor analysis repeatedly generates groups of attributes based on how the surveyed variables are correlated and how many factors to retain. Based on these results, the analyst attempts to name these factors. Note that the factors to be uncovered depend critically on the attributes that are rated by the respondents. Thus, we must be fairly complete in the specification of relevant attributes in order to generate reliable results. Toward that goal, we conducted a first survey to enumerate all relevant attributes, followed by a second survey to collect data for uncovering data quality dimensions.

2.2 First Survey: Generation of Data Quality Attributes

Literature review, brainstorming, and a field study were used in the first survey to generate a fairly complete set of data quality attributes. Our literature review and brainstorming sessions revealed a list of data quality attributes, as shown in Figure 1.

Accessibility
Accuracy
Adaptability
Availability
Breadth
Compatibility
Completeness
Content
Convenience
Correctness
Cost
Credibility
Critical
Data Exchange
Dependability
Depth
Ease Maintenance
Ease of Access
Ease of Update
Ease of Use
Flexibility
Format
Habit
Importance
Integrity
Interpretability
Manipulability
Preciseness
Redundancy
Relevance
Reliability
Reputation
Timeliness
Understandable
Variety
Well-documented

Figure 1 An Initial List of Data Quality Attributes

Since the dimensions of data quality resulted from factor analysis depended, to a large extent, on the attributes that would be discovered in the first survey, we decided that: (1) the number of subjects for the first study should be as large as possible; (2) we should be able to have individual contact with the subjects in order to fully understand their answers; and (3) the subjects should be data consumers with diverse perspectives.

Toward that goal, we interviewed and administered the first survey over the phone to respondents currently working in the industry. In parallel, we conducted the survey at the MIT Sloan School of Management. During the personal interviews with industry respondents, not only were attributes generated but also the attribute's meaning in the interviewee's mind discussed. More than one hundred Sloan MBA's students participated in the self-administered survey. They came from a wide range of industries, and had an average age of more than 30.

As shown in Appendix A, the first survey included two sections for eliciting data quality attributes. The first section was used to elicit the respondents' first reaction to data quality, similar to brainstorming. They were simply asked to list those attributes which first come to mind (in addition to timeliness, accuracy, availability, and interpretability) when they think of data quality. In the second section, the remaining attributes shown in Figure 1 were given to "spark" any additional attributes.

This process resulted in over 170 unique responses, as shown in Figure 2. Only ten attributes were mentioned by more than half of the participants. These results further support the use of factor analysis for uncovering the actual underlying quality dimensions.

2.3 Second Survey: Collecting Data for Uncovering Dimensions

The list of attributes shown in Figure 2 was used to develop the second survey questionnaire. Since we had recorded all unique responses in their original format, there was some degree of duplication, such as "parsimony" vs. "parsimoniousness." When this occurred, we kept the one that was cited most often and eliminated the others. A questionnaire was developed based on the resulting attributes. The question format for factor analysis is simple. This simplicity lends itself to larger response rates and a survey that is more understandable to a larger number of respondents.

Pre-Test

Because of the simplicity of the survey itself, the questionnaire requires only a small number of people to be sampled. Therefore, we solicited eleven respondents: three industry executives, four professionals, two professors, and two MBA students. No major changes were made in the format of the survey as a result of the pre-test. The most significant content change was the elimination of those attributes (or phrases) which a majority of respondents did not understand or did not see any relation between the attributes and data quality. Based on the results from the pre-test, our final second survey questionnaire included 118 data quality questions (i.e., 118 variables for factor analysis), as shown in Appendix B.

Ability to be Joined With
Ability to Download
Ability to Identify Errors
Ability to Upload
Access by Competition
Accessibility
Accuracy
Adaptability
Adequate Detail
Adequate Volume
Aesthetism
Age
Aggregatability
Alterability
Amount of Data
Authority
Availability
Believability
Breadth of Data
Brevity
Certified Data
Clarity
Clarity of Origin
Clear Responsibility over Data
Compactness
Compatability
Competitive Edge
Completeness
Comprehensiveness
Compressibility
Concise
Conciseness
Confidentiality
Conformity
Consistency
Content
Context
Continuity
Convenience
Correctness
Corruption
Cost
Cost of Accuracy
Cost of Collection
Creativity
Critical
Customability
Data Hierarchy
Data Improves Efficiency
Data Overload
Definability
Dependability
Depth of Data
Detail
Detailed Source
Dispersed
Dynamic
Ease of Access
Ease of Comparison
Ease of Correlation
Ease of Data Exchange
Ease of Distinguishing Updated Files
Ease of Maintenance
Ease of Retrieval
Ease of Understanding
Ease of Update
Ease of Use
Easy to Change
Easy to Question
Efficiency
Endurance
Enlightening
Ergonomy
Error-Free
Expandability
Expense
Extendability
Extensibility
Extent
Finalization
Flawlessness
Flexibility
Form of Presentation
Format
Format Integrity
Friendliness
Generality
Habit
Historical Compatibility
Importance
Inconsistencies
Integration
Integrity
Interactive
Interesting
Level of Abstraction
Level of Standardization
Localized
Logically Connected
Manageability
Manipulable
Measurable
Medium
Meets Requirements
Minimality
Modularity
Narrowly Defined
Normality
Novelty
Objectivity
Optimality
Orderliness
Origin
Parsimony
Parsimoniousness
Partitionability
Past Experience
Pedigree
Personalized
Pertinent
Portability
Preciseness
Precision
Proprietary Nature
Purpose
Quantity
Rationality
Redundancy
Regularity of Format
Relevance
Reliability
Repetitive
Reproduceability
Reputation
Resolution of Graphics
Responsibility
Retrievability
Revealing
Reviewability
Rigidity
Robustness
Scope of Info
Secrecy
Security
Self-Correcting
Semantic Interpretation
Semantics
Size
Specificity
Speed
Stability
Storage
Synchronization
Timeliness
Time-indepenency
Translatable
Transportability
Unambiguity
Unbiased
Understandability
Understandable
Uniqueness
Unorganized
Up-to-Dateness
Usable
Usefulness
User Friendly
Valid
Value
Variability
Variety
Verifiable
Volatility
Well-Documented
Well-Presented

Figure 2 Data Quality Attributes Generated from the First Survey

Survey Target Population

We chose to survey the MIT Sloan alumni who reside in the U.S. They consist of individuals in a variety of industries, departments, and management levels, thus satisfying the requirement that our population sample should consist of a wide range of data consumers with different perspectives. We also hoped that the alumni would be more responsive to the questionnaire survey.

The total number of alumni up until 1989 in the United States was 3215. Of this population, we randomly selected 1500, a little less than 50%, individuals. Our survey was mailed along with a cover letter to explain the nature of the study, the time to complete the survey (less than 20 minutes), and its criticality. We gave respondents a six week cut-off period to respond to the survey if they were to be entered into the data set. Most of the alumni received the surveys at their home address. In order to assure a successful survey, we also sent out all the survey questionnaires via first-class mail. As a result, follow-up calls were not needed due to the high response rate by the end of the third week (20%).

3. Data Analysis of the Second Survey Responses

This section presents the overall descriptive statistics of our sample, the specifics of the factor analysis used, and the details of the resulting data quality dimensions.

3.1 Descriptive Statistics

Of the 1500 surveys mailed, 16 were returned because they were undeliverable. Of the remaining 1484, 355 viable surveys were returned by our six week cut-off response date. Surveys with significant missing values or surveys returned by academics were not considered viable. Thus, they were eliminated from our analysis. This represented an effective response rate of 23.92 percent, which is more than sufficient for our purpose.

The responses were spread fairly evenly over industry. Specifically, there was about 28% from the service, 33% from manufacturing, 19% from finance, and the remaining cited "Others." The finance, marketing/sales, and operations departments evenly makeup 40% of the respondents. There were a relatively large number of respondents who circled "Other." Frequently, these respondents were upper level managers, such as presidents or CEO's, or consultants.

What follows are only the highlights of the descriptive statistics for all 118 variables.

Missing Responses: There does not appear to be any attributes (or phrases) that were particularly unclear or hard to answer. While none of the variables had 355 responses, only four had less than 342 responses. The exceptions were *quality of resolution, time independent, robust* and *critical* with 329, 334 , 338 and 333 responses respectively. In addition, there does not appear to be any significant pattern to the missing responses.

Variable Ranges: On our scale where ***1 was extremely important and 9 not important,*** almost every variable had a minimum value of 1 and a maximum value of 9. The exceptions were *accuracy, reliability, level of detail* and *easy identification of errors.* Thus for the majority of variables, there were respondents who felt it was an extremely important attribute, and respondents who felt it was not important at all. *Accuracy* and *reliability* had the smallest range with values ranging from 1 to 7; *level of detail* and *easy identification of errors* went from 1 to 8.

Variable Means: 99 of the variables had means less than or equal to 5. That is, most of the variables surveyed were considered to be important data attributes. The two variables with means less than 2 were *accuracy* and *correct,* with means of 1.771 and 1.816 respectively. *Time independent* had the highest mean of 6.772. Thus, this variable is one of the least important variable in the survey.

3.2 Factor Analysis Specifics and Results

The data quality dimensions were uncovered using factor analysis on the 355 survey responses. All analysis was performed using SYSTAT Version 5.1 for the Macintosh.

Factor Method: We used the multiple principal components method, a variant of factor analysis, on the variable correlation matrix to group variables by factor. We then used the VARIMAX rotation method to clarify the grouping pattern represented by the original principal component dimensions. We chose to use principal components analysis, as opposed to the common factor model, for the following reasons:

> "Principal Components is a reproducible procedure in accounting for common variance in a set of associated variables", whereas "the common factor model does not produce exact factor scores." "Common Factor scores also have to be estimated and there is no

> requirement that the estimated scores be uncorrelated across factors. In principal components, however, uncorrelated component scores are guaranteed in the model" "The components model is less susceptible to misinterpretation, since it entails linear combinations of actual variables." (Green, 1988 #391)

In short, we chose principal component analysis because the results are reproducible, less susceptible to misinterpretation, and factor scores will be uncorrelated across factors.

Factored Matrix: We chose to analyze the correlation matrix instead of the covariance matrix. The resulting component loadings from the correlation matrix represent the correlation of each original variable with each component. Whereas, the covariance loadings represent the covariance of the original variable with each component. Thus, the correlation component loadings are believed by most researchers to be more intuitive measures of variable and factor association.

Convergence Criteria: As specified by the SYSTAT, the convergence criteria for stopping the analysis is either 25 iterations or a tolerance level, which is defined as "the amount of variance an original variable shares with all other variables," of .001 (Hair, 1987 #392). In our case, we reached the tolerance level before 25 iterations.

Limiting the Number of Computed Components: SYSTAT offers two methods for limiting the number of computed components. One can either directly specify the number of desired factors or specify a minimum eigenvalue. A priori specification of the number of components was not an option for our analysis because there is no underlying theory which specifies how many dimensions one would expect to find. Thus, we applied the eigenvalue method and chose to limit the number of components using the "eigenvalue greater than 1" rule.

The "eigenvalue greater than 1" rule makes intuitive sense because it assures that each factor explains at least as much variance as a truly independent variable would explain. In our case, we have 118 variables. Thus, if they were all independent, we would get 118 components. Each would explain 1/118 or .85% of the total variance where 1 is the eigenvalue of the component and 118 is equal to the number of components.

By using the "eigenvalue greater than 1" rule, we limit the number of principle components to the number of variables with eigenvalues greater than 1. Thus, any component after this cutoff number of components explains a smaller amount of variance than an independent variable would explain and does not aid in understanding the factor structure. On the other hand, if one were to set the maximum eigenvalue to be greater than 1, one runs the risk of eliminating possible valuable dimensions. Since our research goal is to uncover new data quality dimensions without eliminating any potential dimensions, the "eigenvalue greater than 1" choice is correct.

Rotation Method: The original principal component solution was rotated using the VARIMAX rotation scheme. It orthogonally rotates the independent components or factors to generate factor loadings which are either close to 1 or 0, making the subsequent assignment of variables to factors as self-evident as possible.

Assignment of Variables to Components: Our resulting components consist of those variables whose rotated component loadings were greater than .5. That is, a variable was assigned to a particular component if the correlation between the component and the variable was at least .5. Although this approach may appear simplistic, it is quite rigorous (Hair, 1987 #392).

3.3 Naming the Dimensions

The initial principal component analysis generated 29 components which explained 73.909 percent of the total variance in the data. Nine components were eliminated based on the following criteria: (1) a .5 loading cut-off point, (2) importance of the component as the respondents rated, and (3) the interpretability of the component. The remaining 20 dimensions explained 59.296% of the total variance, as shown in Table 1.

These dimensions are named as follows:

(1) Believability
(2) Value Added
(3) Relevancy
(4) Accuracy
(5) Interpretability
(6) Ease of Understanding
(7) Accessibility
(8) Objectivity
(9) Timeliness
(10) Completeness
(11) Traceability
(12) Reputation
(13) Representational Consistency
(14) Cost Effectiveness
(15) Ease of Operation
(16) Variety of Data & Data Sources
(17) Conciseness
(18) Access Security
(19) Appropriate Amount of Data
(20) Flexibility

Table 1 Complete list of dimensions (DIM), their adjectives, the component loading (CL), and the % of variance (% VAR) explained by the dimension

DIM	ADJECTIVES	CL	% VAR	DIM	ADJECTIVE	CL	% VAR
1	Believability	0.76	1.408	12			1.801
2			1.991		Reputation of Source	0.78	
	Competitive Edge	0.74			Data Reputation	0.73	
	Adds Value	0.72		13			3.079
3			2.867		Same Format	0.70	
	Applicable	0.74			Consistently Represented	0.66	
	Relevant	0.64			Consistently Formatted	0.57	
	Interesting	0.58			Compatable w/Previous Data	0.57	
	Usable	0.53		14			2.676
4			5.361		Cost of Collection	0.83	
	Certified Error Free	0.78			Cost of Accuracy	0.78	
	Error Free	0.78			Cost Effectiveness	0.71	
	Accurate	0.73		15			7.315
	Correct	0.71			Easily Joined	0.75	
	Flawless	0.66			Easily Integrated	0.71	
	Reliable	0.60			Easily Download/Upload	0.67	
	Easy Identification of Errors	0.58			Easily Aggregated	0.65	
	Integrity	0.54			Easily Customized	0.59	
	Precise	0.51			Easily Updated	0.56	
5	Interpretable	0.64	1.881		Easily Changed	0.56	
6			2.911		Manipulable	0.53	
	Easily Understood	0.70			Used for Multiple Purposes	0.53	
	Clear	0.65			Easily Reproduced	0.53	
	Readable	0.56		16	Variety of Data and Sources	0.68	1.449
7			3.971	17			6.544
	Retrievable	0.68			Well-Presented	0.81	
	Accessible	0.66			Form of Presentation	0.72	
	Easily Accessed	0.58			Concise	0.71	
	Speed of Access	0.57			Well-Organized	0.71	
	Available	0.56			Format of Data	0.69	
	Up-To-Date	0.56			Well-Formatted	0.68	
	Easily Retrieved	0.52			Compactly Represented	0.66	
8			1.777		Aesthetically Pleasing	0.63	
	Unbiased	0.76		18			2.741
	Objective	0.71			No Access	0.77	
9	Age of Data	0.58	1.494		Proprietary	0.75	
10			3.451		Access Can Be Restricted	0.63	
	Breadth of Information	0.85			Secure	0.60	
	Depth of Information	0.81		19	Amount of Data	0.75	1.610
	Scope of Information	0.79		20			2.360
11			2.609		Adaptable	0.58	
	Well Documented	0.72			Flexible	0.56	
	Verifiable	0.64			Extendable	0.53	
	Easily Traced	0.56			Expandable	0.51	
					Total % of Variance		59.296

Our results indicate that the hypothesized dimension *accuracy* is much more complex than previously realized. In fact, a large number of our dimensions can be thought of as relating to accuracy. Yet they were definitely distinct to our set of respondents. These accuracy related dimensions include:

- Accuracy
- Traceability
- Reputation

At first glance, it appears that data consumers view accuracy as one of the most important dimension of data quality. To assure themselves of the accuracy of the data, they also use three additional constructs to defined data quality:

- The definitive knowledge that the data contains no errors
- The ability to trace/verify/audit the data so as to confirm accuracy
- The reputation of the data and data source

Thus, we conclude that *accuracy,* is indeed, extremely important to data quality, and that *accuracy* means much more than the percent of errors in the data. It means that the entire data trail, from initial entry to final results, can be actually traced and examined by the data consumer.

3.4 Elaborating on the Dimensions

We now discuss the dimensions in more detail in order of their importance as dictated by their corresponding means. A dimension mean was computed by forming a new variable for each dimension and calculating the mean response. The new variable consisted of the average of the individual's responses to all of the adjectives with a loading of .5 or greater on the component or dimension. For example, the dimension *ease of understanding* consisted of the three adjectives *easily understood, readable* and *clear*. For each individual response a new variable was created with a value equal to the average of the individual's response to the two adjectives. The overall mean for the dimension is then the average of all responses for this new variable. We now elaborate on the 20 dimensions.

Dimension 1:	**Believability**	**Mean: 2.707**	**CI: 2.505 - 2.909**
Adjective List:	**Believable**		

Believability was a dimension represented by a single adjective. This dimension had a 95% confidence interval which excluded 3, thus we can say that the mean is statistically less than 3. We cannot, however, say that *believability* is the most important dimension, since its confidence interval overlappped with the next three dimensions. *Believable* had a component loading of .76 on this dimension as shown in Table 1, and near zero loadings on the remaining dimensions. No other adjectives loaded clearly on this dimension. Thus, it is argued that *believability* stands on its own and can be interpreted solely by the meaning of the word itself.

Dimension 2:	**Value Added**	**Mean: 2.830**	**CI: 2.647 - 3.013**
Adjective List:	**Data Gives You a Competitive Edge** **Data Adds Value to Your Operations**		

This dimension addresses some of the more qualitative issues of data quality. That is, what benefits do the data consumer obtain from the data itself. As the adjectives in this dimension suggest, the overall value added by the data is very important to data consumers. The only other adjective which loaded relatively high on this dimension, albeit with a value less than .5, was *data improves efficiency*, with a loading of .43. While the mean is not statistically less than 3, *value added* is still one of the most important data quality dimensions since it has a relatively high mean.

Dimension 3:	**Relevancy**	**Mean: 2.951**	**CI: 2.824 - 3.078**
Adjective List:	**Applicable, Relevant, Interesting, Useable**		

The four adjectives listed above all loaded clearly on this dimension. In addition, *important* and *revealing*, also loaded on this dimension with values of .40 and .34 respectively. Thus, this dimension, similar to *value added*, is another more qualitative data dimension, i.e. whether the data user feels that the data is applicable and helpful to the problem or situation at hand.

Dimension 4:	**Accuracy**	**Mean: 3.046**	**CI: 2.856 - 3.236**
Adjective List:	**Data is Certified Error-Free, Error Free, Accurate, Correct, Flawless, Reliable, Errors Can Be Easily Identified, The Integrity of the Data, Precise**		

This dimension confirms the existence of the dimension *accuracy*. All of the adjectives clearly loaded on this dimension since they had near zero loadings on all other components. The particular grouping of these adjectives shows that the concepts of accuracy and error knowledge are closely intertwined. That is, the existence of adjectives such as *errors can be easily identified* and *data is certified error free* combined with adjectives such as *accurate* and *correct* reinforces the point that data users need confirmation that the data is indeed accurate, and that perhaps accuracy, in the eyes of data consumers, can be achieved by making sure that errors can be easily found, instead of eliminating all errors entirely.

The mean of this dimension was also relatively high, and was not statistically different from either of the means for *believability, value added* and *relevancy*. Thus these four dimensions can all be considered to be equally important to data consumers. Other adjectives which seemed to load on this dimenion, although with loadings lower than .5, were *dependable* (.33) and *complete* (.47). Thus one could consider that *accuracy* is an important part of defining dependable data, and that *accuracy* may not only concern errors, but also completeness.

Dimension 5:	**Interpretability**	**Mean: 3.198**	**CI: 3.025 - 3.371**
Adjective List:	**Interpretable**		

This dimension also contained a single adjective which clearly loaded only on it with a value of .64. The other adjectives which also could be considered to load on this dimension were *easily questioned,* with a loading of .34, *familiar* with a loading of .43, and *revealing* with a loading of .46. These adjectives, however, had similar, yet smaller, loadings on other dimensions, so their association was not absolute. We can speculate that perhaps *interpretability of data* concerns whether the data is both understandable and useful to the data consumer. That is, the data is either familiar or revealing, and thus its importance or application is clear. Thus this dimension does seem to be different than simply *ease of understanding,* since the data must not only be easy to understand, but also easy to place in a useful context.

Dimension 6:	**Ease of Understanding**	**Mean: 3.217**	**CI: 3.068 - 3.366**
Adjective List:	**Easily Understood, Clear, Readable**		

Easily Understood loaded higher on this dimension than *clear* or *readable*. The loadings were .70, .65, and .56 respectively, but all three had low loadings on all other dimensions. These adjectives reinforce the interpretation of this dimension as *ease of understanding* instead of simply *understandability*. It appears that *understandable* as defined by this adjectives may have less to do with an in-depth understanding of the data, and more to do with a first glance or more cursory inspection of the data. In addition, it is significant that *easily understood* loaded on this dimension, and not the adjective *understandable*. Thus, we argue that this dimension captures, perhaps, the simplicity of the data presentation more than the understandability of the actual data itself.

Dimension 7: **Accessibility** **Mean: 3.470** **CI: 3.319 - 3.621**
Adjective List: **Accessible, Retrievable, Easily Accessed, Easily Retrieved, Speed of Access Available, Up-To-Date**

This dimension was less concrete than the previous ones, mainly due to the loading of *up-to-date* on the dimension. It should be noted, however, that *up-to-date* had the smallest loading, .56, and had a few relatively high loadings on other dimensions, thus making its association with this dimension less concrete. Second tier adjectives which loaded on this dimension included *transportable /portable* (.40) and *convienent* (.38). Both of these adjectives loaded highest on this dimension, but did also have other similar loadings. We chose to interpret this dimension as the *accessibility* of the data to the data consumer. The dimensional mean of *accessibility* is statistically higher than 3. Thus, it did not rank in the extreme end of importance to data consumers. However, it is still relatively high both on the dimension list and in terms of its mean.

Dimension 8: **Objectivity** **Mean: 3.577** **CI: 3.395 - 3.759**
Adjective List: **Unbiased, Objective**

Both *unbiased* and *objective* loaded specifically on this dimension with high loadings of .76 and .71 respectively. No other adjectives loaded on this dimension. The mean was statistically smaller than 4, but also greater than 3. From these results, *objectivity* was fairly important to data consumers.

Dimension 9: **Timeliness** **Mean: 3.640** **CI: 3.426 - 3.854**
Adjective List: **Age of Data**

This dimension was also a single adjective dimension, with a relatively low loading value of .58. The only second tier adjective which clearly loaded on this dimension was *"it is easy to tell if the data is updated"*, with a loading of .42. We label this dimension as *timliness* because the second tier adjective does lead to the interpretation of the dimension as perhaps *"currency"* or *"up-to-dateness"*.

Dimension 10: **Completeness** **Mean: 3.880** **CI: 3.704 - 4.056**
Adjective List: **The Breadth of Information Contained in the Data**
The Depth of Information Contained in the Data
The Scope of Information Contained in the Data

We chose to call this dimension *completeness* because the three phrases were related to the importance of the overall content of the data as the user sees it. This dimension had high loadings on the adjectives of .85, .81, and .79, respectively. The only second tier adjective which clearly loaded on this dimension was *"specific."*(.42).

Dimension 11: Traceability Mean: 3.965 CI: 3.786 - 4.144
Adjective List: Well-Documented, Easily Traced, Verifiable

In examining the loadings, we see that *well documented* has a loading of .72, easily traced a loading of .56, and verifiable a loading of .64. No other adjectives with the exception of *auditable* (.43) loaded clearly on this dimension. With this in mind, we chose to label the dimension as *traceability* since it deals with whether the data trail can be traced, followed, or verified by data consumers.

Dimension 12: Reputation Mean: 4.039 CI: 3.832 - 4.246
Adjective List: The Reputation of the Data Source, The Reputation of the Data

This dimension evidentially concerns the trust or regard the data consumer has in the actual data source and data content. Both phrases loaded highly on this dimension, with near zero loadings on all other dimensions. It was also true that no other adjectives loaded clearly on the *reputation* dimension. This dimension may represent a way to easily assure data consumers of the trustworthiness of the data, and hence increase perceived data quality, without expensive or extensive overhauls of the current Information Systems. In addition, this dimension is similar to an issue in consumer quality -- the difference between perceived quality and actual quality. It might be interesting for IS departments to compare the reputation of their data and data sources to their actual quality levels. In some instances, correcting misunderstandings may provide data quality benefits with little cost.

Dimension 13: Representational Consistency Mean: 4.216 CI: 4.040 - 4.392
Adjective List: Data Is Continuously Presented In Same Format, Consistently Represented Consistently Formatted, Data is Compatible with Previous Data

On the surface, it does not seem surprising that the above adjectives were "grouped together" into a single dimension. However, these adjectives were positioned in very different places in the survey, and this type of grouping did not always occur when expected. Thus it appears that data consistency is definitely a construct which consumers use to think about or evaluate quality data. The adjective *"you have used the data before"* also loaded relatively high on this dimension with a loading of .35. Thus it appears that concistency has possibly two benefits, one is that the data is compatible with previous data, and the other is that the data is familiar. In addition, this dimension had a mean statistically less than 5, thus it is relatively important to data consumers.

Dimension 14: Cost Effectiveness Mean: 4.246 CI: 4.051 - 4.441
Adjective List: Cost of Data Accuracy, Cost of Data Collection, Cost Effective

This dimension was the only dimension that address the cost aspect of data quality. Therefore, it was not surprising that it is an important dimension. Further support is found by looking at the the mean of 4.246, whose 95 percent confidence interval had a low end statistically less than 5. In looking at the component loadings, *cost of collection* (.83), *cost of accuracy* (.78) and *cost effectiveness* (.71), we see that they were rather high, and each adjective had near zero loadings on all other dimensions. Again this supports *cost effectiveness* as being a unique dimension.

Dimension 15: Ease of Operation Mean: 4.281 CI: 4.125 - 4.437
Adjective List: Easily Joined, Easily Changed, Easily Updated, EasilyDownloaded/Uploaded Data Can be Used for Multiple Purposes, Manipulable, Easily Aggregated, Easily Reproduced, Data Can Be Easily Integrated, Easily Customized

Many of the adjectives that were associated with ease were "loaded" on a single dimension which we called *ease of operation*. However, this dimension is not just a result of the position in the questionnaire or expression of the adjectives. This is apparent because those adjectives which did not relate to the manipulability of the data per se, did not load on this dimension, even though they appeared in the same section. In addition, *manipulable* and *multiple purposes* both did load on this dimension. Thus it is apparent that *ease of operation* is a valid data quality dimension. *Ease of operation* was one of the dimensions that highlights the operational data quality issues.

The mean of this dimension is statistically less than 5 which indicates that it is an important dimension in the eyes of the data consumer.

Dimension 16: Variety of Data & Data Sources Mean: 4.712 CI: 4.476 - 4.948
Adjective List: You Have a Variety of Data and Data Sources

Variety of Data & Data Sources is also an isolated adjective. As analysis shows, *variety of data & data sources* had low loadings on all other dimensions and the only second tier adjective which loaded highly on this dimension was *"the source of the data is clear"* (.43). Thus this dimension represents the existence of a choice of data sources available to data consumers, and these sources are evident to the user so that he/she can chose judiciously between them.

Dimension 17: Concise **Mean: 4.753 CI: 4.585 - 4.921**
Adjective List: Well-Presented, Concise, Compactly Represented, Well-Organized
Aesthetically Pleasing, Form of Presentation, Well-Formatted
Format of the Data

The adjectives listed above loaded relatively high on this dimension. In looking at these adjectives, they, like *easy of understanding,* seem to deal with the cursory inspection of the data. Second tier adjectives which loaded predominantly on this dimension included "*you have little extraneous data present*" (.44) and "*the data is not overwhelming*" (.45). These adjectives both reinforce the interpretation of this dimension as *concise* and not merely the presentability of the data.

Dimension 18: Access Security **Mean: 4.922 CI: 4.704 - 5.140**
Adjective List: Data Cannot be Accessed by Competitors, Data Is of a Proprietary Nature
Access To Data Can Be Restricted, Secure

This is also a new dimension for data quality for the reasons described below. All of these adjectives loaded high on this dimension while loading low on the other dimensions. Specially, the loading were .77, .75, .63, and.60 respectively. No other adjectives clearly loaded on this dimension. Since they all were in relation to security of the system/applications from others, it became obvious to call it *access security.* This also reinforces the claim that data quality is more than just the content of the data itself. Data Quality includes how one can interact with the systems/applications.

Dimension 19: Appropriate Amount of Data **Mean: 5.009 CI: 4.785 - 5.233**
Adjective List: The Amount of Data

Amount of data was another dimension represented by a single adjective, yet the loading was very high on this dimension (.75), and the adjective had near zero loadings on all other dimensions. The only other adjective which appeared to load on this dimension was "you have little extraneous data present," which loaded with a value of .36. These two adjectives together can be interpretted as a desire for the appropriate amount of data that can be used to effectively address the data consumer's needs. This represents conciseness, not in presentation form, but in actual data content. It is interesting to note, however, that the mean of this dimension was statistically less important than that of concise.

Dimension 20: Flexibility **Mean: 5.340 CI: 5.166 - 5.514**
Adjective List: Adaptable, Flexible, Extendable, Expandable

This dimension has a 95% confidence which excludes 5. Thus, it is on the low end of importance of the dimensions presented. All four the adjectives loaded on the dimension with respective loadings of .58, .56, .53, and .51. No other adjectives loaded clearly on the dimension. Thus, based on the above, the dimension was labelled *flexibility*.

4. Summary and Future Directions

We are actively conducting research along the following directions: What kinds of information technologies can be developed to certify existing corporate data; to certify external sources of data; and to provide data auditability? What kinds of operations management techniques can be applied to help develop a research foundation for data quality management? How should data originators, data distributors, and data consumers manage data quality problems differently, or should they not? What is the relationship between data quality and the corresponding data attributes in the context of risk management? These inquiries will help develop a body of knowledge for data quality management -- an increasingly critical issue facing Corporate America for the decade to come.

Grouping the Dimensions of Data Quality

Value Added (2) Cost Effectiveness (14)			
Ease of Understanding(6)	Believability (1)	Relevancy (3)	Timeliness (9)
Interpretability (5) Representational Consistency (13) Conciseness (17)	Accuracy (4) Objectivity (8) Completeness (10) Traceability (11) Reputation (12) Variety of Data & Data Sources (16) Access Security (18)	Appropriate Amount of Data (19)	Accessibility (7) Ease of Operation (15) Flexibility (20)

APPENDIX A: DATA QUALITY FIELD SURVEY

Position Prior to Attending Sloan:
(Circle One)

Finance Marketing
Operations Personnel
IT Other____________

Industry you worked in the previous job : ____________________

When you think of data quality, what dimensions **other than** timeliness, accuracy, availability, and interpretability come to mind? Please list as many as possible!

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

PLEASE FILL OUT THIS SIDE BEFORE TURNING OVER. THANK YOU!!

- -

- -

(Side Two)

The following is a list of dimensions developed for data quality:

Completeness
Relevance
Ease of Update
Integrity
Well-documented
Dependability
Ease of Access
Understandable

Flexibility
Reputation
Ease Maintenance
Breadth
Habit
Manipulability
Convenience
Credibility

Adaptability
Compatibility
Format
Depth
Variety
Preciseness
Accessibility
Importance

Reliability
Ease of Use
Cost
Correctness
Content
Redundancy
Data Exchange
Critical

After reviewing this list, do any other dimensions come to mind?

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

THANK YOU!

Appendix B: Second Survey Questionnaire

Thank you for participating in this study. All responses will be held in strictest confidence.

Industry: ______________________ Job Title: __________________________

Department: Finance Marketing/Sales Operations Human Resources
Accounting Information Systems Planning Other ______________

The following is a list of adjectives and phrases which describes corporate data. When answering the questions, please think about the internal data such as sales, production, financial, and employee data that you work with or use to make decisions in your job.

We apologize for the tedious nature of the survey. Although the questions may seem repetitive, your response to each question is critical to the success of the study. Please give us the first response that comes to mind and try to use the FULL scale range available.

Section I: How important is it to you that your data is:

	Extremely Important				Important			Not Important At All	
Accurate	1	2	3	4	5	6	7	8	9
Believable	1	2	3	4	5	6	7	8	9
Complete	1	2	3	4	5	6	7	8	9
Concise	1	2	3	4	5	6	7	8	9
Verifiable	1	2	3	4	5	6	7	8	9
Well-Documented	1	2	3	4	5	6	7	8	9
Understandable	1	2	3	4	5	6	7	8	9
Well-Presented	1	2	3	4	5	6	7	8	9
Up-To-Date	1	2	3	4	5	6	7	8	9
Accessible	1	2	3	4	5	6	7	8	9
Adaptable	1	2	3	4	5	6	7	8	9
Aesthetically Pleasing	1	2	3	4	5	6	7	8	9
Compactly Represented	1	2	3	4	5	6	7	8	9
Important	1	2	3	4	5	6	7	8	9
Consistently Formatted	1	2	3	4	5	6	7	8	9
Dependable	1	2	3	4	5	6	7	8	9
Retrievable	1	2	3	4	5	6	7	8	9
Manipulable	1	2	3	4	5	6	7	8	9
Objective	1	2	3	4	5	6	7	8	9
Useable	1	2	3	4	5	6	7	8	9
Well-Organized	1	2	3	4	5	6	7	8	9
Transportable/Portable	1	2	3	4	5	6	7	8	9
Unambiguous	1	2	3	4	5	6	7	8	9
Correct	1	2	3	4	5	6	7	8	9

Please try to use the FULL scale range available.

Section I (continued): How important is it to you that your data is:

	Extremely Important				Important			Not Important At All	
Relevant	1	2	3	4	5	6	7	8	9
Flexible	1	2	3	4	5	6	7	8	9
Flawless	1	2	3	4	5	6	7	8	9
Comprehensive	1	2	3	4	5	6	7	8	9
Consistently Represented	1	2	3	4	5	6	7	8	9
Interesting	1	2	3	4	5	6	7	8	9
Unbiased	1	2	3	4	5	6	7	8	9
Familiar	1	2	3	4	5	6	7	8	9
Interpretable	1	2	3	4	5	6	7	8	9
Applicable	1	2	3	4	5	6	7	8	9
Robust	1	2	3	4	5	6	7	8	9
Available	1	2	3	4	5	6	7	8	9
Revealing	1	2	3	4	5	6	7	8	9
Reviewable	1	2	3	4	5	6	7	8	9
Expandable	1	2	3	4	5	6	7	8	9
Time Independent	1	2	3	4	5	6	7	8	9
Error-Free	1	2	3	4	5	6	7	8	9
Efficient	1	2	3	4	5	6	7	8	9
User-Friendly	1	2	3	4	5	6	7	8	9
Specific	1	2	3	4	5	6	7	8	9
Well-Formatted	1	2	3	4	5	6	7	8	9
Reliable	1	2	3	4	5	6	7	8	9
Convenient	1	2	3	4	5	6	7	8	9
Extendable	1	2	3	4	5	6	7	8	9
Critical	1	2	3	4	5	6	7	8	9
Well-Defined	1	2	3	4	5	6	7	8	9
Reusable	1	2	3	4	5	6	7	8	9
Clear	1	2	3	4	5	6	7	8	9
Cost Effective	1	2	3	4	5	6	7	8	9
Auditable	1	2	3	4	5	6	7	8	9
Precise	1	2	3	4	5	6	7	8	9
Readable	1	2	3	4	5	6	7	8	9

Section II: How important is it to you that your data can be:

	Extremely Important				Important			Not Important At All	
Easily Aggregated	1	2	3	4	5	6	7	8	9
Easily Accessed	1	2	3	4	5	6	7	8	9
Easily Compared to Past Data	1	2	3	4	5	6	7	8	9
Easily Changed	1	2	3	4	5	6	7	8	9
Easily Questioned	1	2	3	4	5	6	7	8	9
Easily Downloaded/Uploaded	1	2	3	4	5	6	7	8	9
Easily Joined With Other Data	1	2	3	4	5	6	7	8	9
Easily Updated	1	2	3	4	5	6	7	8	9
Easily Understood	1	2	3	4	5	6	7	8	9
Easily Maintained	1	2	3	4	5	6	7	8	9
Easily Retrieved	1	2	3	4	5	6	7	8	9
Easily Customized	1	2	3	4	5	6	7	8	9
Easily Reproduced	1	2	3	4	5	6	7	8	9
Easily Traced	1	2	3	4	5	6	7	8	9
Easily Sorted	1	2	3	4	5	6	7	8	9

Section III: How important are the following to you?

	Extremely Important				Important			Not Important At All	
Data is certified error-free.	1	2	3	4	5	6	7	8	9
Data improves efficiency.	1	2	3	4	5	6	7	8	9
Data gives you a competitive edge.	1	2	3	4	5	6	7	8	9
Data cannot be accessed by competitors.	1	2	3	4	5	6	7	8	9
Data is in finalized form.	1	2	3	4	5	6	7	8	9
Data contains no redundancy.	1	2	3	4	5	6	7	8	9
Data is of proprietary nature	1	2	3	4	5	6	7	8	9
Data can be personalized.	1	2	3	4	5	6	7	8	9
Data is not easily corrupted.	1	2	3	4	5	6	7	8	9
Data meets all of your requirements.	1	2	3	4	5	6	7	8	9
Data adds value to your operations.	1	2	3	4	5	6	7	8	9
Data is continuously collected.	1	2	3	4	5	6	7	8	9
Data continuously presented in same format.	1	2	3	4	5	6	7	8	9
Data is compatible with previous data.	1	2	3	4	5	6	7	8	9
Data is not over whelming.	1	2	3	4	5	6	7	8	9

Section III (Continued): How important are the following to you?

	Extremely Important				Important			Not Important At All	
Data can be easily integrated.	1	2	3	4	5	6	7	8	9
Data can be used for multiple purposes.	1	2	3	4	5	6	7	8	9
Data is secure.	1	2	3	4	5	6	7	8	9

Section IV: How important are the following to you?

	Extremely Important				Important			Not Important At All	
The source of the data is clear.	1	2	3	4	5	6	7	8	9
Errors can be easily identified.	1	2	3	4	5	6	7	8	9
The cost of data collection.	1	2	3	4	5	6	7	8	9
The cost of data accuracy.	1	2	3	4	5	6	7	8	9
The form of presentation.	1	2	3	4	5	6	7	8	9
The format of the data.	1	2	3	4	5	6	7	8	9
The scope of information contained in data.	1	2	3	4	5	6	7	8	9
The depth of information contained in data.	1	2	3	4	5	6	7	8	9
The breadth of information contained in data.	1	2	3	4	5	6	7	8	9
Quality of resolution.	1	2	3	4	5	6	7	8	9
The storage medium.	1	2	3	4	5	6	7	8	9
The reputation of the data source.	1	2	3	4	5	6	7	8	9
The reputation of the data.	1	2	3	4	5	6	7	8	9
The age of the data.	1	2	3	4	5	6	7	8	9
The amount of data.	1	2	3	4	5	6	7	8	9
You have used the data before.	1	2	3	4	5	6	7	8	9
Someone has clear responsibility for data.	1	2	3	4	5	6	7	8	9
The data entry process is self-correcting.	1	2	3	4	5	6	7	8	9
The speed of access to data.	1	2	3	4	5	6	7	8	9
The speed of operations performed on data.	1	2	3	4	5	6	7	8	9
The amount and type of storage required.	1	2	3	4	5	6	7	8	9
You have little extraneous data present.	1	2	3	4	5	6	7	8	9
You have a variety of data and data sources.	1	2	3	4	5	6	7	8	9
You have optimal data for your purpose.	1	2	3	4	5	6	7	8	9
The integrity of the data.	1	2	3	4	5	6	7	8	9
It is easy to tell if the data is updated.	1	2	3	4	5	6	7	8	9
Easy to exchange data with others.	1	2	3	4	5	6	7	8	9
Access to data can be restricted.	1	2	3	4	5	6	7	8	9

Thank you for your time and effort in completing this survey.

5. References

[1] Ballou, D. P. & Tayi, G. K. (1989). Methodology for Allocating Resources for Data Quality Enhancement. *Communications of the ACM, 32,* pp. 320-329.

[2] Bodner, G. (1975). Reliability Modeling of Internal Control Systems. *Accounting Review,* 50(October), pp. 747-757.

[3] Cash, J. I. & Konsynski, B. R. (1985). IS redraws competitive boundaries. *Harvard Business Review,* 63(2), pp. 134-142.

[4] Clemens, E. (1988). McKesson drug company: a case study of economost, a strategic information system. *Journal of Management Information Systems,* 5(1), pp. 141-149.

[5] Date, C. J. (1990). *An Introduction to Database Systems* (5th ed.). Reading, MA: Addison-Wesley.

[6] Garvin, D. A. (1987). Competing on the eight dimensions of quality. *Harvard Business Review,* (November-December), pp. 101-109.

[7] Garvin, D. A. (1988). *Managing Quality-The Strategic and Competitive Edge* (1 ed.). New York: The Free Press.

[8] Goodhue, D. L., Quillard, J. A., & Rockart, J. F. (1988). Managing The Data Resources: A Contingency Perspective. *MIS Quarterly,* 12(3), pp. 373-392.

[9] Hansen, J. V. (1983). Audit Considerations in Distributed Processing Systems. *Communications of the ACM,* 26(August), pp. 562-569.

[10] Hansen, M. & Wang, Y. R. (1990). *Managing Data Quality: A Critical Issue for the Decade to Come.* Composite Information Systems Laboratory, MIT December 1990.

[11] Henderson, J. C. (1989). *Building and sustaining partnership between line and I/S managers.* (CISR WP #195) 1989.

[12] Ives, B. & Learmonth, G. P. (1984). The information system as a competitive weapon. *Communications of the ACM, 27,* pp. 1193-1201.

[13] Keen, P. G. W. (1986). *Competing In Time: Using Telecommunications For Competitive Advantage* (1 ed.). Ballinger.

[14] Laudon, K. C. (1986). Data Quality and Due Process in Large Interorganizational Record Systems. *Communications of the ACM,* 29(1), pp. 4-11.

[15] Lehmann, D. R. (1989). *Market Research and Analysis*, Irwin. Boston: Mass .

[16] Lindgren, B. (1991). Getting data with integrity. *Enterprise,* (winter), pp. 30-34.

[17] Madnick, S., Osborn, C., & Wang, Y. R. (1990). Motivating Strategic Alliances for Composite Information Systems: the case of a major regional hospital. *Journal of Management Information Systems,* 6(4), pp. 99-117.

[18] Madnick, S. E. & Wang, Y. R. (1988). Evolution towards strategic applications of databases through composite information systems. *Journal of Management Information Systems,* 5(2), pp. 5-22.

[19] Maier, D. (1983). *The Theory of Relational Databases* (1st ed.). Rockville, MD: Computer Science Press.

[20] McFarlan, F. W. (1984). Information technology changes the way you compete. *Harvard Business Review,* 62(2), pp. 98-105.

[21] Morton, S. M. (1989). *Management in the 1990s: Research Program Final Report.* 1989.

[22] Rockart, J. F. & Short, J. E. (1989). IT in the 1990s: Managing Organizational Interdependence. *Sloan Management Review, Sloan School of Management, MIT,* 30(2), pp. 7-17.

Dimension 17: Concise **Mean: 4.753** **CI: 4.585 - 4.921**
Adjective List: **Well-Presented, Concise, Compactly Represented, Well-Organized**
Aesthetically Pleasing, Form of Presentation, Well-Formatted
Format of the Data

The adjectives listed above loaded relatively high on this dimension. In looking at these adjectives, they, like *easy of understanding,* seem to deal with the cursory inspection of the data. Second tier adjectives which loaded predominantly on this dimension included "*you have little extraneous data present*" (.44) and "*the data is not overwhelming*" (.45). These adjectives both reinforce the interpretation of this dimension as *concise* and not merely the presentability of the data.

Dimension 18: Access Security **Mean: 4.922** **CI: 4.704 - 5.140**
Adjective List: **Data Cannot be Accessed by Competitors, Data Is of a Proprietary Nature**
Access To Data Can Be Restricted, Secure

This is also a new dimension for data quality for the reasons described below. All of these adjectives loaded high on this dimension while loading low on the other dimensions. Specially, the loading were .77, .75, .63, and.60 respectively. No other adjectives clearly loaded on this dimension. Since they all were in relation to security of the system/applications from others, it became obvious to call it *access security.* This also reinforces the claim that data quality is more than just the content of the data itself. Data Quality includes how one can interact with the systems/applications.

Dimension 19: Appropriate Amount of Data **Mean: 5.009** **CI: 4.785 - 5.233**
Adjective List: **The Amount of Data**

Amount of data was another dimension represented by a single adjective, yet the loading was very high on this dimension (.75), and the adjective had near zero loadings on all other dimensions. The only other adjective which appeared to load on this dimension was "you have little extraneous data present," which loaded with a value of .36. These two adjectives together can be interpretted as a desire for the appropriate amount of data that can be used to effectively address the data consumer's needs. This represents conciseness, not in presentation form, but in actual data content. It is interesting to note, however, that the mean of this dimension was statistically less important than that of concise.

Dimension 20: Flexibility **Mean: 5.340** **CI: 5.166 - 5.514**
Adjective List: **Adaptable, Flexible, Extendable, Expandable**

This dimension has a 95% confidence which excludes 5. Thus, it is on the low end of importance of the dimensions presented. All four the adjectives loaded on the dimension with respective loadings of .58, .56, .53, and .51. No other adjectives loaded clearly on the dimension. Thus, based on the above, the dimension was labelled *flexibility*.

3.5 Grouping the Dimensions

Table 2 Grouping the Dimensions of Data Quality

Value Added (2) Cost Effectiveness (14)			
Ease of Understanding(6)	Believability (1)	Relevancy (3)	Timeliness (9)
Interpretability (5) Representational Consistency (13) Conciseness (17)	Accuracy (4) Objectivity (8) Completeness (10) Traceability (11) Reputation (12) Variety of Data & Data Sources (16) Access Security (18)	Appropriate Amount of Data (19)	Accessibility (7) Ease of Operation (15) Flexibility (20)

4. Summary and Future Directions

We are actively conducting research along the following directions: What kinds of information technologies can be developed to certify existing corporate data; to certify external sources of data; and to provide data auditability? What kinds of operations management techniques can be applied to help develop a research foundation for data quality management? How should data originators, data distributors, and data consumers manage data quality problems differently, or should they not? What is the relationship between data quality and the corresponding data attributes in the context of risk management? These inquiries will help develop a body of knowledge for data quality management -- an increasingly critical issue facing Corporate America for the decade to come.