

Measure, Analyze, and Improve IQ

*T*o manage effectively, one must measure and analyze. To measure, however, one must first define what to measure. Defining what Information Quality (IQ) means, therefore, is crucial in managing information as a product. In the process of defining what IQ means, the participants will engage in problem identification and problem solving in the context of their own organizational setting. They must identify the necessary organizational processes and technical solutions for managing the information product. Toward that goal, the Information Product Manager (IPM) [10] must develop the corresponding IQ metrics, upon defining IQ dimensions, to measure and analyze the quality of the information product and improve it accordingly. This chapter addresses these issues.

MEASURE IQ

Invariably, companies concerned about IQ pose three basic questions:

How good is the quality of information in my databases, data warehouses, or customer information systems?

How does the quality of my information compare to that of others in my industry? Is there a set of benchmarks that we can use as a basis of comparison?

Is there a usable, single, aggregate IQ measure, similar to stock market indicators, such as the Dow Jones Industrial Average?

To answer these questions, the IPM must develop a suitable set of metrics to perform the necessary measurements. The requirement to measure is inextricably intertwined with the needs to analyze and improve IQ. Unfortunately, there is no one universal, invariant set of metrics that can be used by everyone. There is no “one size fits all” set of metrics or, for that matter, no one universal number that measures IQ. An aggregate, weighted function can be developed, but this will be specific to one company and reflect subjective assignment of the weights.

In developing IQ metrics, it is important to recognize the many factors that should be considered. All too often, however, firms develop their IQ metrics based on their intuition or previous experience. As a result, sustainable long-term improvement is not achieved. This is particularly unfortunate after millions of dollars have been invested in implementing a software system for improving IQ, and tens of million more dollars are required to maintain and modify the system. A negative consequence is that these firms become cynical when their IQ initiatives fail to achieve the planned objectives.

Based on the cumulative body of IQ research and practice, we propose three complementary classes of metrics that the IPM must recognize and use when assessing IQ metric needs. Specifically, the IPM must obtain an accurate assessment of the perception of IQ from the key organizational functional units. The IPM must have a set of application-independent baseline measures of the quality of the information — the necessary conditions for IQ. Lastly, the IPM must develop a set of application-dependent measures that reflect specific requirements and business rules of the organization. In other words, the IPM must have three classes of metrics:

Metrics that measure an individual’s subjective assessment of IQ (how good do people in our company think the quality of our information is)

Metrics that measure IQ quality along quantifiable, objective variables that are application independent (how complete, consistent, correct, and up-to-date the information in our customer information system is)

Metrics that measure IQ quality along quantifiable, objective variables that are application dependent (how many clients have exposure to the Asian finan-

cial crisis that our risk management system cannot estimate because of poor quality information)

Used in combination, metrics from each of these classes provide fundamental information that goes beyond the static IQ assessment to the dynamic and continuous evaluation and improvement of information quality. Each class measures something different. The metrics that provide subjective evaluations relate to one individual's perception of the quality of information. The application-independent metrics transcend specific applications and are context independent. The application-dependent metrics have meaning and relevance to a specific application and are context dependent [8].

By approaching the development of IQ metrics from the perspective of these three classes, the IPM can diagnose the current status of IQ, develop the required metrics, link these metrics to the organization's goals and objectives, and conduct cost-benefit analyses allowing management to make informed decisions regarding initiatives to improve the quality of information.

Subjective IQ Metrics

Subjective IQ metrics measure an individual's subjective assessment of IQ. That is, how good do people in our company think the quality of our information is. Two types of subjective IQ measures can be established. The first type measures the dimensional IQ, while the second measures the level of IQ knowledge in the firm.

Dimensional IQ Assessment

Based on the 16 IQ dimensions [11] presented in the previous chapter, we can generate a set of questions (questionnaire) to determine the perception of the state of IQ in an organization. Such a questionnaire has been developed based on the cumulated research conducted at MIT's TDQM program [3]. A complete copy of this questionnaire is shown in the appendix at the end of this chapter. Each question is rated using a Likert-type scale on a scale from 0 to 10 where 0 indicates "not at all" and 10 "completely." A sample question to assess the dimension on completeness is: "This information is sufficiently complete for our needs."

This questionnaire has been used effectively in both public and private sectors. For example, IS managers in one investment firm thought they had perfect IQ (in terms of accuracy) in their organizational databases. However, following their completion of the questionnaire, they found deficiencies such as,

additional information about information sources was needed so that information consumers could assess the reputation and believability of the information

information downloaded to servers from the mainframe was not sufficiently timely for some information consumers' tasks

the currencies (\$, £, or ¥) and units (thousands or millions) of financial information from different servers were implicit so information consumers could not always interpret and understand this information correctly

The questionnaire can be used to measure perceived IQ. It can be used as a diagnostic tool to evaluate the quality of information from a much broader perspective than the limited perspective of information accuracy only. Information obtained during this diagnostic phase provides the motivation to develop methods for improving the quality of information as perceived by information consumers. Such methods could include users. Aside from its diagnostic uses, the framework and associated questionnaire are valuable aids when used as checklists during information requirements analysis.

In addition to the assessment of IQ, a number of questions are included to assess the degree to which a company has in place mechanisms to ensure the quality of information. These mechanisms include TQM programs, IQ software tools, and an IQ administration function. A complete copy of this questionnaire is shown in Section 3 of the Appendix. Each question is rated using a Likert-type scale using values from 1 to 10 where 1 indicates "very small extent" and 10 "very large extent." A sample question is: "In this company, there are people whose primary job is to assure the quality of information." Analysis of the degree to which these mechanisms have been instituted helps the firm to plan, execute, and monitor the firm's IQ program.

IQ Knowledge Assessment

The previous discussion focused on the assessment by the information consumer of the current quality of information, and the mechanisms that the firm deploys to assure high-quality information. To maintain a viable total IQ program in the firm, knowledge of the system in place to manage IQ must also be assessed.

Three aspects of knowledge pertinent for shaping organizational IQ capabilities have been identified [5]. This perspective explicitly includes aspects of knowledge that probe underlying reasons and axiomatic assumptions behind the work practice in organizations. Specifically, *IQ-related "know-what" knowledge* is the accumulated understanding of the activities and procedures involved in producing, storing, and utilizing information. *IQ-related "know-how" knowledge* is the accumulated skills for applying routine procedures to known IQ problems. *IQ-related "know-why" knowledge* is the ability to analyze and discover previously unknown IQ problems or solutions. This type of knowledge is gained from experience and understanding of the objectives and cause-effect relationships underlying the activities involved in collecting, storing, and utilizing information. These three types of knowledge apply to each of the three information manufacturing processes: information production, storage, and use.

To assess the state of these three types of IQ knowledge in the firm, a separate questionnaire has been developed. The questions are intended to assess the level of IQ knowledge in the firm. Each question is rated using a Likert-type scale using values from 1 to 10 where 1 indicates “very small extent” and 10 “very large extent.” A sample question is: “I know which group collects this information.”

The responses from the IQ knowledge questionnaire can be analyzed to identify problem areas and to develop corresponding solutions. Repeated applications of these questionnaires in companies in different industries will eventually lead to the establishment of a representative data set. These standards provide IQ benchmarks for individual organizations in different industries. The results of these questionnaires will be used to analyze and improve the IQ.

Objective, Application-Independent Metrics

Objective, application-independent metrics measure IQ quality along quantifiable, objective variables; for example, how complete, consistent, correct, and up-to-date the information in our customer information system is. These metrics are based on established theory for controlling the quality of information entering the system. Systems for which these controls are not in place at the time of information acquisition can still use the measures to assess the degree to which the existing information meets the standard.

Most database systems have been designed from a systems perspective. Mechanisms such as integrity constraints and normalization theories [1, 2, 7] used to maintain the integrity and consistency of information are necessary, but not sufficient, to attain quality information as demanded by information users.

Dr. Edgar F. Codd proposed five integrity rules that must be followed by any true relational database management system. Although developed specifically for the relational model, the integrity rules are very useful in many contexts ranging from the relational database systems to network database systems to flat files. Indeed, many of the corporate databases are just beginning to be migrated to the relational database environment such as IBM’s DB2 from the decade-old CICS hierarchical database management systems. Moreover, much of current-day information is stored in either spreadsheets or groupware such as Lotus Notes. Ensuring the quality of information in these databases requires a methodological approach, and integrity rules proposed by Codd present such an approach.

Even in the relational environment, many present-day relational databases are not Integrity-compliant for many reasons.

- The specific relational database implementation does not have an integrity facility to enforce integrity.

- Although the integrity facility is available, the DBA does not provide the specifics.

- While the edit checks are programmed into the database management system, the user by necessity often overrides the rules when under time pressure.

As the business environment changes, so do the business rules that must be enforced on the underlying data.

Simply put, Codd's integrity rules ensure data meet the specifications demanded by the designer and the user. An IQ tool developed based on Codd's integrity rules, therefore, offers a rigorous method to define, measure, analyze, and improve the quality of information. Below we recap the five kinds of integrity Codd proposed.

Domain Integrity — All the values of a field must be of the same domain.

Column Integrity — specifies the set of acceptable values for the column.

Entity Integrity — No component of a primary key is allowed to have a missing value of any type. No foreign key is allowed to have a missing and inapplicable value.

Referential Integrity — For each distinct foreign key in a relational database, there must exist in the database an equal value of a primary key from the same domain. If the foreign key is composite, those components that are themselves foreign keys must exist in the database as components of at least one primary key value drawn from the same domain.

User Defined Integrity — This captures business rules and company regulations and operations that should be reflected in the database. User-defined constraints are used not only to ensure the state of the database is valid but also to trigger specific actions when specified conditions arise in the database.

Application-Dependent IQ Metrics

Application-dependent metrics measure IQ quality along quantifiable, objective variables that are domain specific and require domain experts' participation. Some application-dependent metrics are relatively intuitive and easy to develop whereas others may be very involved. Below we illustrate these problems and opportunities in a financial company case study.

Financial Company is a leading investment bank with extensive domestic and international operations. The nature of its business requires the company to have an accurate and up-to-date representation of each customer's risk profile. Investments at inappropriate customer risk levels cause major customer dissatisfaction and potential indemnification to customers for losses. The company was not poised to leverage customer account information in its global operations. For example, customers with sufficient credit across accounts could not trade or borrow on their full balance. Tracking customer balances for individual and multiple accounts, closing all accounts of an investor because of criminal activities, and ensuring an accurate investor risk profile also could not be accomplished without significant, error-prone human intervention. All of these presented the company with potentially huge problems and many opportunities.

The company established a few information process measures or controls. For example, there were no controls to ensure that customer risk profiles were updated on a regular basis. The account creation process was not standardized or inspected. Consequently, no metrics were established to measure how many accounts were created on time, and whether customer information in those accounts was updated. Managing its customer account information as a product would provide Financial Company with better risk management and customer service — two critical success factors for companies in the financial industry.

In the client account database, for example, the following IQ metrics could be applied:

- the percentage of incorrect client address zip code found in a randomly selected customer accounts (*inaccuracy*)
- an indicator of when client account information was last updated (*timeliness* or *currency* for database marketing and regulatory purposes)
- the percentage of nonexistent accounts or the number of accounts with missing values in the industry-code field (*incompleteness*)
- the number of records that violate referential integrity (*consistency*)

At a more complex level, there are business rules that need to be observed. For example, the total risk exposure of a client should not exceed a certain limit. This exposure needs to be monitored for clients who have many accounts. Conversely, a client who has a very conservative position in one account should be allowed to execute riskier transactions in another account. For these business rules to work, however, the firm needs to develop a proper linking method to link the accounts.

There are also information-manufacturing-oriented IQ metrics. In the client account system, for example, the firm needs to track

- which department made most of the updates in the system last week
- how many unauthorized accesses have been attempted (*security*)
- who collected the raw data for a client account (*credibility*)

Another example is Financial Company's mission-critical Capital Markets System. Consider its account reconciliation which is conceptually similar to balancing a checkbook. For a checkbook, one may

- record all checks for completeness
- inspect checks for accuracy
- adjust the beginning balance with checks issued during the period for internal consistency with the end balance
- compare the checkbook balance with bank balance for external validity

monitor large checks and control possible overspending in certain categories
improve the check balancing process as necessary

Account reconciliation in the Capital Markets System, however, is more complicated than balancing a checkbook. It involves thousands of transactions drawn on hundreds of accounts stored in computer and manual systems around the world. Each account has a one-to-one relationship with a ledger. An item in a ledger, which is the bank's record of payments made and expected, could be a payment of £6 million to an Exxon office in London. These items need to be matched with items in statements from hundreds of other banks in which this international bank has accounts. Tens of billions of U.S. dollars are reconciled each day.

Many IQ problems arise during reconciliation. For example, a statement may not be received in time for reconciliation or could have items with wrong account numbers. The sources of poor IQ could be recording, transmission, or simply missing information. Some example causes are

An operator delayed sending a statement or entered wrong data.

Instead of paying £6 million to Exxon, London directly, three separate items were created: pay £1 million to Ford, £2 million to GM, and £3 million to Chrysler.

For operational and regulatory reasons, it is critical to minimize the number of items, called open items, not reconciled. Based on results from this research, an IQ monitor is being developed in the bank to minimize open items. Open items that do not match are prioritized according to the number of days since initially identified as an open item.

To verify internal consistency, an IQ monitor adds all the candidate items in statements for an account (e.g., Exxon) to the beginning balance in a ledger to see if they match the ending balance. If the result is inconsistent, the candidate items are listed as open items.

To verify cross consistency, an IQ monitor matches items in a ledger with those in a statement (and vice versa). Candidate item-pairs are compared to see if the account number, amount, debit/credit, date, and so on are consistent. Those with no match are listed as open items. Efforts are made to reconcile these problems by, for example, calling the party who produced the items. Credibility of information source is also important in reconciling open items; some banks are notorious for poor quality information in their statements, and therefore, items from those statements are often checked first during reconciliation.

In addition to the above relatively obvious IQ issues, there are business-oriented rules. For example, it is important to keep track of items that are more than \$10,000 and remain open for more than 10 days.

ANALYZE IQ

We have presented three classes of IQ measures and argued that when used in combination, these measures provide fundamental information that goes beyond the static IQ assessment to the dynamic and continuous evaluation and improvement of information quality. To perform the necessary IQ analysis efficiently and effectively, however, it would be useful to have some computer-based tools to facilitate the analysis. In this section, we present such a software implementation and illustrate how these tools can be applied.

IQ Assessment (IQA)

The *IQ Assessment*[™] (IQA) Survey consists of three sections. Section 1 collects the characteristics of the information source being assessed. The subject is required to answer four questions in this section.

Question 4 elicits the role the subject plays in activities involved in this database. The four primary roles are information producer, information consumer, information manufacturer, and manager of these positions. Our experience shows that often a subject will declare multiples roles of involvement. In this case, the IQA Survey administrator should ask the subject to focus on the primary role played for purposes of completing the questionnaire. The department in which a subject is employed often determines the primary role. For example, a subject who works in the MIS department is most likely to be an information manufacturer, an information vendor representative such as one from Reuters or Dun & Bradstreet is an information producer; and a marketing manager whose is responsible for a direct market campaign is most likely to be an information consumer.

The IQA software instrument is developed to collect the survey data electronically, following the principle that the quality of information will be high if entered by subjects themselves (the very information source). We illustrate the electronic questionnaire below using question 2. As shown in the appendix at the end of this chapter, the subject would respond to question 2 by selecting the appropriate numerical response from 1 to 10. The response that the subject has chosen will also be displayed to the left of the selection buttons. If the subject wishes, the responses can be typed in directly in the box to the left of the button. Section 2 of the IQA assesses the dimensional quality of the information. It would be completed in the same manner as Section 1 except that the screen contains more than one question. The subject would select the desired numerical rating for each question. Figure 4.1 illustrates the screen for Section 2.

Section 2: Information Quality Assessment

For each statement, indicate the extent to which it is true of this information.
 "This information" refers to the information or database selected by your company
 for reporting on in this information quality questionnaire.

		Not at all	Completely									
1. This information is easy to manipulate to meet our needs.		0	1	2	3	4	5	6	7	8	9	10
2. It is easy to interpret what this information means.	9	0	1	2	3	4	5	6	7	8	9	10
3. This information is consistently presented in the same format.		0	1	2	3	4	5	6	7	8	9	10
4. This information includes all necessary values.		0	1	2	3	4	5	6	7	8	9	10
5. This information is easily retrievable.	10	0	1	2	3	4	5	6	7	8	9	10
6. This information is formatted compactly.	7	0	1	2	3	4	5	6	7	8	9	10
7. This information is protected against unauthorized access.	5	0	1	2	3	4	5	6	7	8	9	10
8. This information is incomplete.	4	0	1	2	3	4	5	6	7	8	9	10
9. This information is not presented consistently.	4	0	1	2	3	4	5	6	7	8	9	10
10. This information has a poor reputation for quality.	9	0	1	2	3	4	5	6	7	8	9	10
11. This information is complete.	1	0	1	2	3	4	5	6	7	8	9	10
12. This information is presented concisely.	0	0	1	2	3	4	5	6	7	8	9	10
13. This information is easy to understand.	4	0	1	2	3	4	5	6	7	8	9	10
14. This information is believable.	2	0	1	2	3	4	5	6	7	8	9	10
15. This information is easy to aggregate.	8	0	1	2	3	4	5	6	7	8	9	10
16. This information is of sufficient volume for our needs.	4	0	1	2	3	4	5	6	7	8	9	10

< Go to Section 1 Goto Section 3 >

Figure 4.1 Sample IQA Screen

(Source: *Cambridge Research Group* [3])

For most of the questions, the subject can simply use the mouse to click a button to indicate the response. Alternatively, the subject can respond to the questions using the keyboard. In that case, the subject will use the tab key to position the cursor in the appropriate box, enter the desired numerical rating, and tab again to move to the next question.

Section 3 of the IQA collects the contextual quality. It is completed in the same manner as Section 2. Note, however, that in Section 3 a response of N/A is permissible because specific questions on context will not apply to all organizations.

As mentioned earlier, the questions are repetitive and some are reverse coded to ensure the validity of the questionnaire. Survey results that violate the validity are excluded from further analysis. Although the IQ Survey software is developed for the subject to enter responses electronically, a subject can complete a hard copy version of the questionnaire instead of using the IQ Survey software. In that case, the administrator can enter the results directly with the software's survey administrator function.

IQA Survey Administration

In preparing for the administration of the IQ Assessment Survey, the survey administrator works with key stakeholders of an IQ project to pick an information system whose information is (mission) critical to the firm and for which IQ improvement is important to the firm. Before administering the IQA survey, the survey administrator should emphasize to the subjects that

The survey is not a test. There is no correct answer to each of the questions. The subject serves as an informed representative in answering the survey questions.

Survey questions are repetitive and some are reverse coded, again, to verify the validity of the results.

It takes about 8 minutes to complete the survey.

There is no need to check previous answers for consistency.

Survey Results Analysis: A Case Study

After the IQA Survey results have been collected and entered into the IQA Survey Master Database, the analysis task can take place. The results of analysis will vary depending on the many contextual factors in an IQ project. Accordingly, expertise in IQ management and familiarity with the project are essential in developing insightful results. Statistical packages such as SPSS, SAS, and Minitab can be applied to facilitate the analysis of the survey results. A proprietary software tool to analyze results of the survey has also been developed [3]. Below, we present a case study illustrating the use of this tool to conduct an analysis.

Appliance Company sells personal computers, home office products, consumer electronics, entertainment software, major appliances and related accessories through its retail stores. The company generate billions of dollars of revenue annually. Prior to the year we investigated the quality of information in the company, net income fell 100 percent compared to the previous year. Revenues reflect the opening of new stores. Earnings were offset by the inability to leverage certain components of its fixed costs, and an increase in interest expense.

To compete more effectively, senior management demanded more accurate, timely, and relevant snapshot reports. As part of a mission-critical thrust to meet the demand, consultants were called upon to assess the IQ of the underlying databases from which these reports are produced. Following the survey administration process as presented above, data points were collected from various functional areas that encompass information producers, information manufacturers, information consumers, and those who are responsible for managing the production of the reports as an information product.

Analysis of the survey results provided answers to the following questions:

Q1: How is Appliance Company's IQ?

Q2: Which IQ dimensions are of high/low quality?

Q3: How do different groups assess high/low quality information?

To answer the first question, an unweighted, aggregate IQ index was computed, yielding a score of 6.77 out of 10. As with many other self-assessment questionnaires, this overall score is merely an index that reflects the participants' views of Appliance's IQ (similar to that of Dow Jones Industrial Average as an index to the U.S. equity market).

The answer to the second question is shown in Table 4.1. Overall, the data from snapshot reports rate high along the intrinsic quality category, which includes the dimensions of believability and reputation. Poor quality shows up in the contextual and representational IQ categories. In order for Appliance Company to improve IQ, it needs to focus on improving completeness of information so as to reflect and plan its business performance. It also needs to improve ease of manipulation so that information can easily be adapted for analysis and other business purposes.

Table 4.1: Overall Information Quality

	High-quality IQ Dimensions*	Low-quality IQ Dimensions**
Overall	H1. Believability H2. Reputation H3. Relevancy	L1. Ease of manipulation L2. Security L3. Appropriate amount of data L4. Completeness

* These IQ dimensions are evaluated as high by the survey participants.

** These IQ dimensions are evaluated as poor by the survey participants.

To answer the third question, the survey results are aggregated by the roles the subjects assumed. The results of analyzing the responses for perceived quality as a function of the role of the respondent are depicted in Figure 4.2. It is evident that information custodians (mostly MIS department) view the information as very timely, but information consumers disagree. Information consumers in all groups view information from snapshot reports as not easy to manipulate for their business purposes, but information custodians disagree [9].

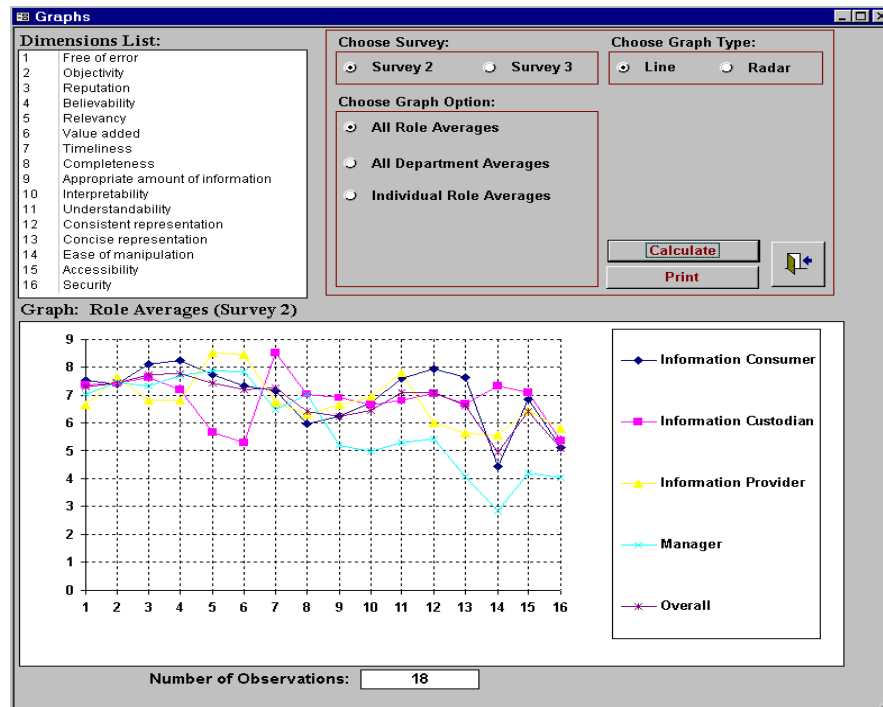


Figure 4.2 IQ Assessment across Roles

(Source: Cambridge Research Group [3])

We have presented the IQA Survey software tool and illustrated, through the Appliance Company case study, an analysis based on the survey results collected from Appliance. This survey, along with interviews, can diagnose critical areas and directions for IQ improvement. Moreover, these qualitative IQ assessment results can be combined with the results of both system-based and application-based metrics. The combination of results from the IQ Survey, system-based metrics, and application-based measurements contribute to the development of a comprehensive picture of a company's overall level of IQ. We present a software implementation below.

Integrity Analyzer™

An outgrowth of research from MIT's Total Data Quality Management (TDQM) research, the *Integrity Analyzer*™ (IA) embeds a TDQM methodology that combines the principles of the TDQM cycle with the principles of integrity constraints in relational databases, as shown in Table 4.2 [4, 6]. Each column of Table 4.2 represents one of the five integrity constraints defined by Codd: domain, entity, referential, column, and user-defined integrity. Domain integrity requires that all values in a column of a table must be drawn from the same domain. Entity integrity requires that every entity (table) must have a primary key consisting of one or more columns. The primary key must be unique and have no missing values. Referential integrity requires that, for each distinct foreign key in a relational database, there must exist in the database an equal value of a primary key from the same domain. Column integrity further restricts the values that can be drawn from the domain for that particular column. In short, column integrity specifies the set of acceptable values for the column. These values can be specified in the form of uniqueness requirement, nonnull requirements, a range of acceptable values, or a list of acceptable values. User-defined integrity specifies additional business rules that column values must meet. These rules often involve conditions dependent on values of other fields. The four actions, *define*, *measure*, *analyze*, and *improve*, for IQ constitute the rows of Table 4.2. These four actions can be applied to achieve domain, entity, referential, column, and user-defined integrity. The cells in Table 4.2 represent the application of an IQ action to a type of integrity.

Table 4.2: A TDQM Methodology for Integrity Analyzer

	Domain Integrity	Entity Integrity	Referential Integrity	Column Integrity	User defined Integrity
Define	Define the domains used in the database. For each column, specify its associated domain.	For each table, specify the primary key, and any candidate keys.	Specify all foreign keys. For each, specify its associated primary key.	For each column, specify the rules for acceptable values.	Specify any business rules not captured in entity, referential, and column integrity.
Measure	Check for violations, i.e., values in a column that are not drawn from the appropriate domain.	Check for violations, i.e., keys that are null or non-unique.	Check for violations, i.e., foreign key values that have no corresponding primary key value.	Check for violations, i.e., column values that have unacceptable values, e.g., out of range.	Check for violations, i.e., record instances that do not satisfy the business rules.
Analyze	Examine the measurement statistics nu-	Examine the measurement statistics nu-	Examine the measurement statistics nu-	Examine the measurement statistics	Examine the measurement statistics numeri-

	merically or graphically.	merically or graphically.	merically or graphically.	numerically or graphically.	cally or graphically.
Improve	View the violation records and change values as appropriate.	View the violation records and change values as appropriate.	View the violation records and change values as appropriate.	View the violation records and change values as appropriate.	View the violation records and change values as appropriate.

We illustrate the IA's functionality by using the example of an international bank that is migrating legacy files to a relational database. This problem typifies a problem that companies face when migrating data from legacy systems or spreadsheet files to a relational database or data warehouse. All too often, the quality of the information in the source files is poor. This state occurs, in part, because integrity constraints were not applied when the data were originally entered. Our example consists of five files that have been migrated: the customer file, the account file, account type file, transactions file, and type of transaction type file. Figure 4.3 shows an entity-relationship diagram for this example.

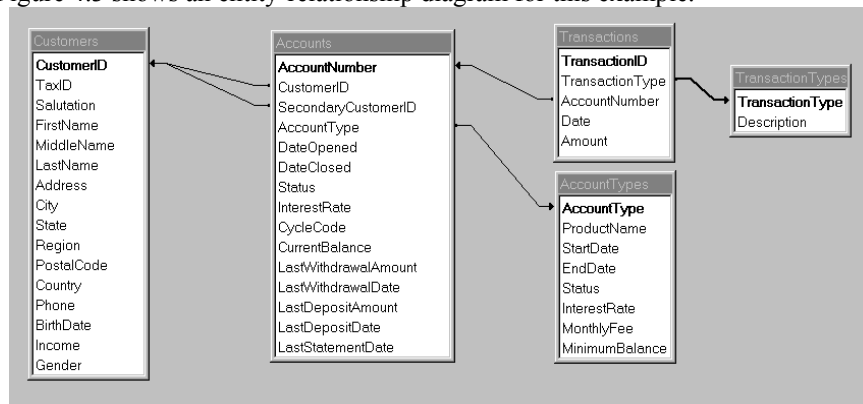


Figure 4.3: An Entity-Relationship Diagram for a Sample Financial Application

(Source: Cambridge Research Group [4])

To illustrate the use of the IA, we first assume that the data have been migrated but an IQ assessment has not been conducted. A user would typically assess

the soundness of the database structure, that is, assess primary and foreign key quality.

the quality of the data in nonkey attributes.
that data conform to the business rules of the firm.

In this example, we focus on the sequence of events and the user-system interface for achieving these objectives.

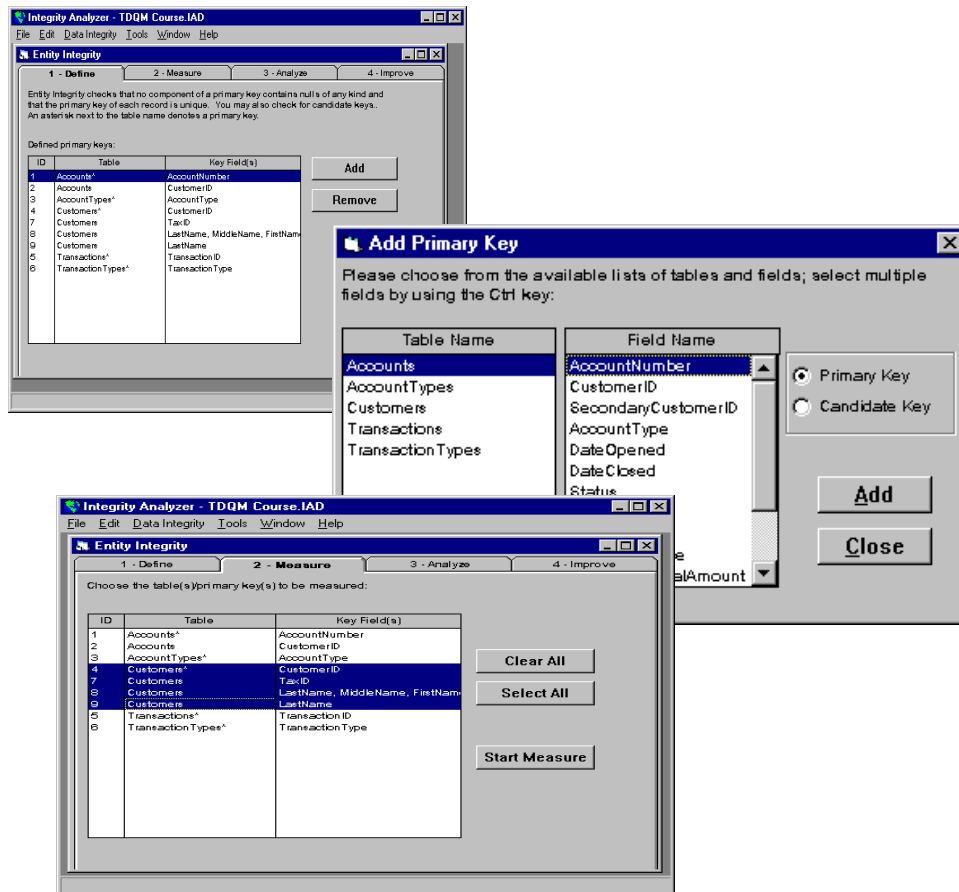
At the start of a session, the IA gives the user the choice of either opening an existing project file or creating a new project file. For a new project file administrative information such as the name and location of the database and a name for the improvement project is obtained. Information in the project file is maintained throughout the life of the project and consists of administrative information on the project, definitions of the structures, and any quality measurements previously made. The user can then select any of the four integrity functions by pulling down the *data integrity* menu and making the desired choice. To perform the frequency check the user would pull down the *tools* menu and select the desired function.

Data Integrity

The entity integrity function checks that all primary and candidate keys for each table are unique and nonnull. The referential integrity function checks that all foreign keys have corresponding primary key values.

To check entity integrity, the IA user selects entity integrity from the data integrity pull-down menu and selects the Define tab. In the Define list box the user selects the fields that are primary and candidate keys for each table. Next the user selects the Measure tab and asks the system to measure the number of violations to entity integrity. After the assessment has been completed, the user can select the Analyze tab and have the violation statistics displayed in numerical, graphical, or report form. Selecting the Improve tab produces a data object that displays the violation instances. The Referential integrity check works in a similar manner.

Exhibits of the screens corresponding to the functions of defining, measuring, analyzing, and improving entity integrity are shown in Figure 4.4



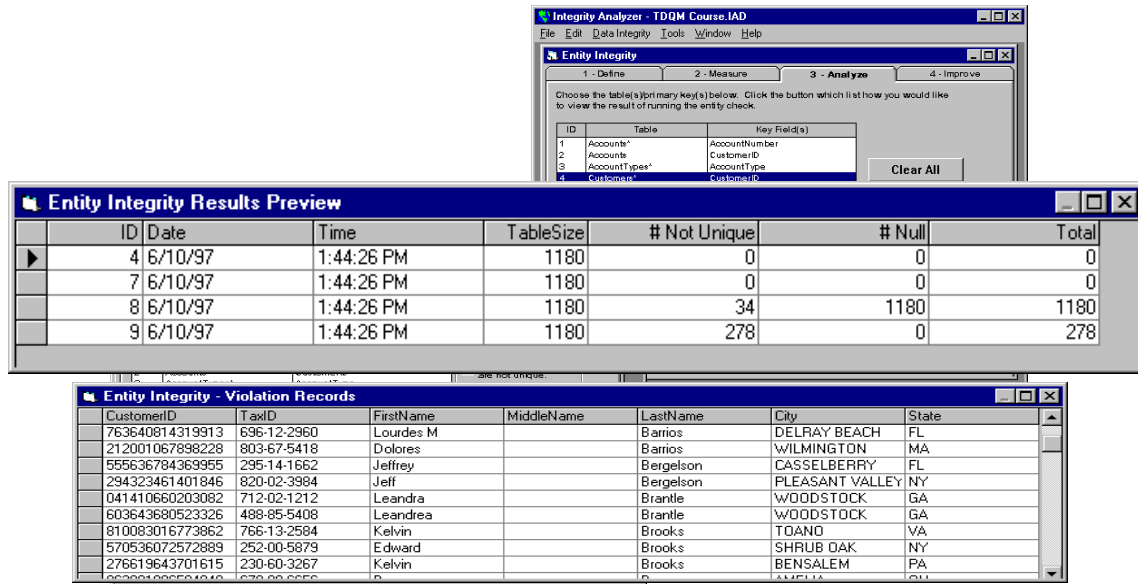
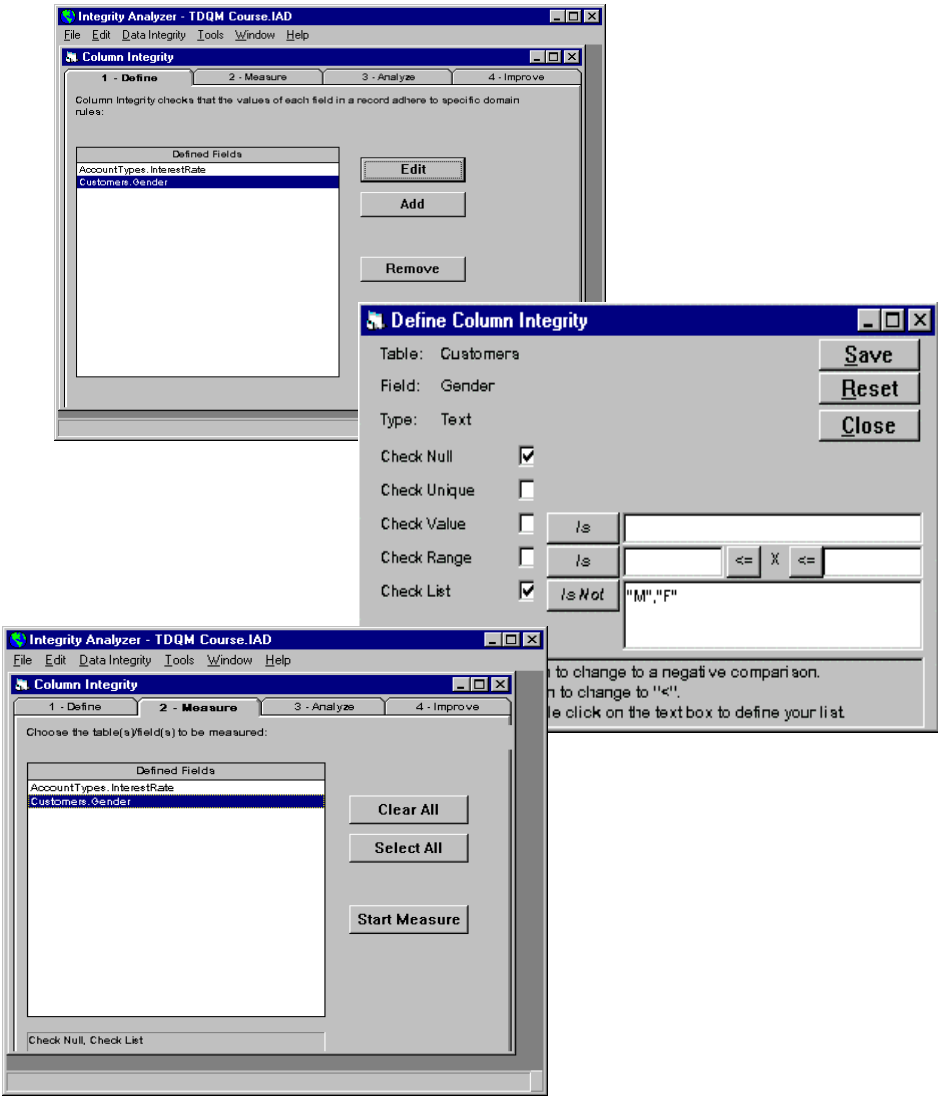


Figure 4.4: Entity Integrity Implementation

(Source: Cambridge Research Group [4])

The example we use to illustrate the features of column integrity is the checking of the values in the gender field of the customer record. The values should be either M or F and not null. After choosing column integrity, the user specifies the table, the column, the type of data, and one or more checks with the Define function. In our example, the user has selected the Customers table, the Gender field whose data type is text, and the “Check Null” and “Check List” options. The appropriate test for the list is constructed by using the condition button adjacent to the Check List selection window and specifying the desired values in the adjacent space provided. In the example shown in Figure 4.5, “Is Not” has been selected and the specific values are “M” and “F.”

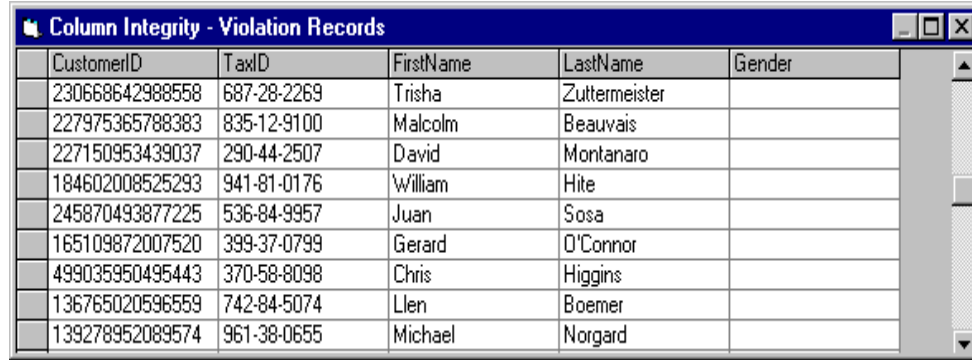
Selecting the Measure tab initiates the execution of the assessment. A number of different reports can be displayed by using the Analyze feature. Selecting the Improve tab will produce a listing of records in violation of the condition (or more precisely meeting the conditions specified which is the same as the specification of the violation being checked). Two conditions are shown in Figure 4.5: records with values other than “M” or “F” and records with null values.



The screenshot displays the Integrity Analyzer - TDQM Course.IAD software interface. The main window is titled "Integrity Analyzer - TDQM Course.IAD" and has a menu bar with "File", "Edit", "Data Integrity", "Tools", "Window", and "Help". The "Data Integrity" menu is open, showing options for "Column Integrity", "Table Integrity", "Field Integrity", and "Record Integrity". The "Column Integrity" window is active, showing a list of defined fields and a list of checks. The "Defined Fields" list includes "AccountTypes", "InterestRate", and "Customers", with "Customers" selected. The "Check(s)" list includes "Check Null" and "Check List", with "Check List" selected. The "View / Edit" button is visible at the bottom of the "Column Integrity" window.

Below the main window, a separate window titled "Column Integrity - Violation Records" is shown, displaying a table of violation records. The table has the following columns: CustomerID, TaxID, FirstName, LastName, City, State, and Gender. The data is as follows:

CustomerID	TaxID	FirstName	LastName	City	State	Gender
223115327891349	895-55-7316		Otero and Pearl Assoc	WEST ROXBURY	MA	C
215703334603744	315-41-0702		Cosmopolitan Proper	PITTSBURGH	PA	C
166517233904626	870-44-6836		Gallagher Realty	NAPLES	FL	C
161985551632639	450-93-8818		Re-Max Affiliates	ELMIRA	NY	C
108563410427518	976-34-5283		Re-Max First Realty	ABINGDON	MD	C
067820199027929	614-03-7433		J.Walsh Real Estate	ALLISON PARK	PA	C
025041470976620	304-02-7425		Jack Conway & Com	KIRKVILLE	NY	C
013505187943770	984-17-5470		BSCH Associates	LEESBURG	FL	C
012982556394228	571-40-4960		Century 21 Cityside	TAMPA	FL	C
196739921647784	999-62-0512	Yinkau	Ho	PUNTA GORDA	FL	X
173093520467254	088-48-8554	P.	Gartin	CHATTANOOGA	TN	X
128192362036103	559-65-0820	Jody	Twombly	VERGENNES	VT	X
113042667641983	272-05-0499	Costa	Georgopoulos	PISCATAWAY	NJ	X
068487203588638	385-66-3517	P.	Goloskie	HAMPTON	VA	X



CustomerID	TaxID	FirstName	LastName	Gender
230668642988558	687-28-2269	Trisha	Zuttermeister	
227975365788383	835-12-9100	Malcolm	Beauvais	
227150953439037	290-44-2507	David	Montanaro	
184602008525293	941-81-0176	William	Hite	
245870493877225	536-84-9957	Juan	Sosa	
165109872007520	399-37-0799	Gerard	O'Connor	
499035950495443	370-58-8098	Chris	Higgins	
136765020596559	742-84-5074	Llen	Boemer	
139278952089574	961-38-0655	Michael	Norgard	

Figure 4.5: Column Integrity Implementation

(Source: Cambridge Research Group [4])

User-defined integrity captures rules that are application dependent. Typically, these would be specific business rules. Examples of such rules are

A student account does not accrue interest.

ATM transactions will incur a fee beginning January 1, 1997.

A senior account holder should be more than 65 years old.

When a customer turns 65 years old, a “senior account” status should be offered.

A customer with an average balance of \$1M should be flagged for new business opportunities.

These business rules range from strict integrity constraints to rules that support marketing and customer support functions. The IA functionality for user-defined integrity is demonstrated below for the first example, a student account does not accrue interest.

Having chosen the User-Defined Integrity function, the user is presented with a display of conditions that have been defined. The user can now edit an existing condition or add a new user-defined rule. If the user selects add, the system displays a Define window that elicits the information necessary to define a condition. The user defined rule is that student accounts do not accrue interest, that is IF Account type = Student, THEN Interest rate = 0.

As in the previous examples, selection of the Measurement function evaluates the database for violation of this rule. Selection of the Analysis function displays the results of the

assessment. Selection of the Improvement function will result in a display of the records that violate the condition.

Frequency Checks

Several additional tools to support IQ analysis are provided with the IA. A particularly useful one is frequency checks, which reports the values and frequencies of each value in a column. Frequency checks is often used in combination with column integrity. Consider the gender field we checked for M and F values with column integrity. The results of integrity checking show many violations. As demonstrated in Figure 4.6, frequency checks can provide information to further analyze these violations.

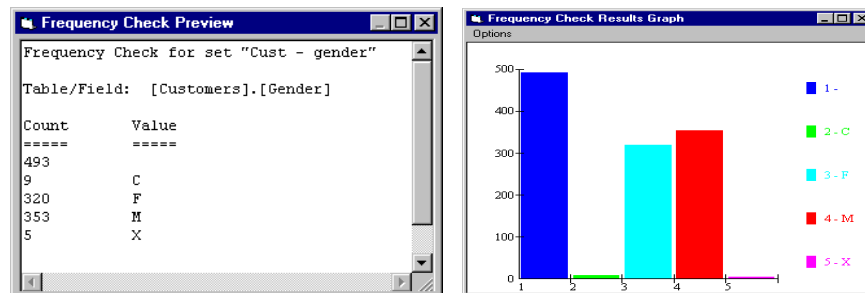


Figure 4.6: Frequency Check

(Source: Cambridge Research Group [4])

As in the examples of previous features, the user can define one or more specific frequent checks or edit a previously defined frequency check specification. By selecting the Measure function, the user initiates a specific frequency check. The user can display results of the assessment by selecting the Analyze option. The results of a frequency check on the gender field of the Customer table are shown in Figure 4.6.

The *Integrity Analyzer* supports both Codd's formally defined integrity constraints and his user-defined constraints. Assessment of domain integrity, column integrity, entity integrity, and referential integrity is invariant across applications. All relational database management systems, ideally, should abide by these rules. All databases, regardless of subject matter, should adhere to the rules represented by these constraints.

The user-defined constraints, however, are application dependent. They vary from application to application, and from industry to industry. Furthermore, these application-

dependent constraints evolve over time. The *Integrity Analyzer* can be customized for a specific company within a specific industry by coding these user-defined rules in the software.

We note that the integrity analyzer is more than simply an implementation of Codd's integrity constraints in relational database software. Unlike the standard, commercial Relational Data Base Management Systems (RDBMS) packages which check for adherence to Codd's constraints when data are entered into the database, the *Integrity Analyzer* is a diagnostic tool which can assess the degree to which existing databases adhere to all the constraints defined by Codd as well as application-dependent user-defined rules. As a diagnostic tool, the analyzer delivers an analysis of the data repository's current quality state and suggests where improvements must be made.

By using the IQA and the integrity analyzer software tools, an analyst can assess the consumer's perceptions of IQ and specific objective states of IQ. These assessments will form the basis for improvement in the overall IQ of the firm.

IMPROVE IQ

It is important that both technical solutions and organizational processes be introduced, disseminated, and institutionalized in the organization over time in order to sustain long-term improvement of organizational information quality. By managing information as a product, an information product management team is established to define what an information product is, and how to manage the information product or information product line over its life cycle. Techniques and methods developed in many disciplines can be applied to the various phases of the TDQM cycle and the information product cycle, for example, statistical process control in production, record tracing and auditing in accounting, and product services in marketing.

Depending on the organizational context, the technical solutions can range from data scrubbing to integrity enforcement to dummy record tracing to data dictionary standardization to complete an information system's overhaul. Similarly, because information quality requirements evolve over time, the organizational processes must capture and monitor the shifting demand of information quality requirements of soundness, usability, usefulness, and dependability. An encoded file that is accurate is sound but not usable if the consumer can not decode it. The information is not useful if does not add value to the consumer's tasks. The IQA instrument can be applied to monitoring the evolving changes of IQ requirements. For example, an organization began its corporatewide IQ initiative. As part of the initiative, more than 30 executives participated in the IQA survey. As shown in Figure 4.7, the results from 28 observations indicate that accessibility is more of a concern than accuracy.

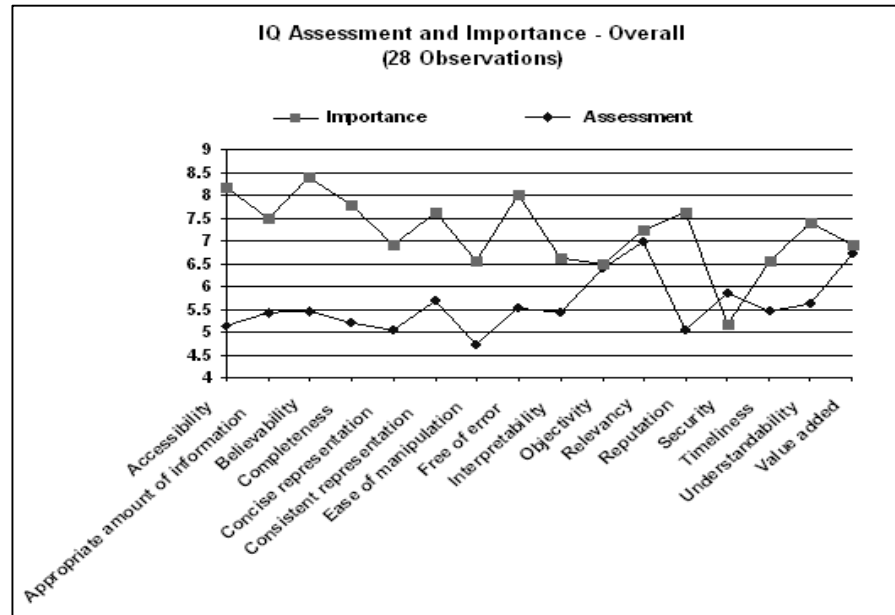


Figure 4.7: IQ Assessment and Importance Ratings

(Source: Cambridge Research Group [3])

A weighted rating, based on the distance between the assessment of a dimension and the ideal score times the importance of the dimension, provides a better indication of how to prioritize the tasks. This is illustrated in Figure 4.8 Note that over time when the accessibility problem has been addressed, an IQA survey of the participants would show another dimension such as timeliness as the primary concern. Furthermore, as with the quality of a physical product, such as a car, the consumer's expectation becomes higher as they become more sophisticated. Information previously considered as accurate, timely, or accessible may not be so as time goes because of a higher expectation.

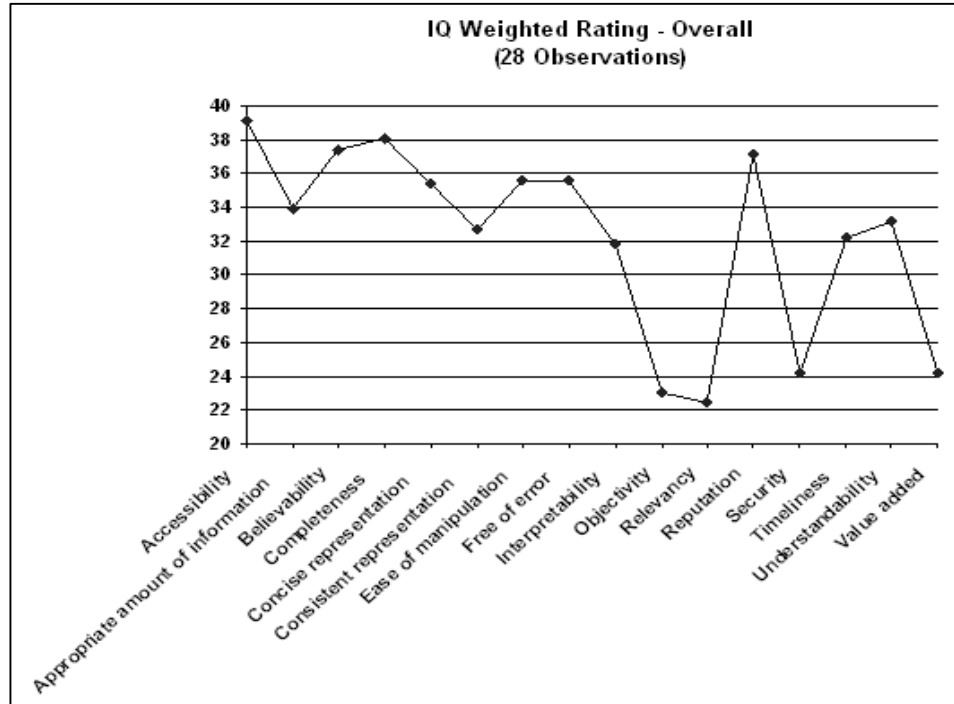


Figure 4.8: Weighted Ratings for Prioritizing IQ Tasks

(Source: Cambridge Research Group [3])

CONCLUSION

We have presented an innovative methodology for managing the quality of organizational information beyond conventional approaches that are limited to the adaptation of TQM techniques to the IQ arena or some specific software tools that scrub data, match names, or reconcile data semantics. Central to our methodology is the concept of managing information as a product. In substantiating the methodology, we developed fundamental concepts such as the TDQM cycle and the information manufacturing system. We also developed a comprehensive set of dimensions to capture the essence of information quality based on a rigorous field studies that premised on the belief that information quality needs to be de-

defined as information that is fit for use by information consumers, as well as a set of ontologically grounded IQ dimensions that established a basis for the rigorous development ultimately of a theory for information quality. With the concept and principles of information quality well defined, we developed tools, methods, and techniques for measuring, analyzing, and improving information quality.

The road to information quality can be bumpy. To facilitate the process, we have presented many cases to illustrate how firms go about launching an information quality initiative, such as introducing, disseminating, and institutionalizing their organizational IQ processes and system solutions. Properly managed, the information product will meet or exceed the expectations of the various stakeholders of the product, and in the long term become part of the organizational culture. Sound, usable, useful, and dependable information is quality information. It is the keystone for organizational knowledge management, the subject that we will present in depth based on IBM's experience and success in developing and deploying their Intellectual Capital Management (ICM) solutions.

APPENDIX: IQA SURVEY

Information Quality Study

This study is being conducted in cooperation with your company and MIT's Total Data Quality Management (TDQM) Research Program. We ask you to assess the quality of your organization's information along many dimensions of quality. These quality dimensions all relate in some way to whether the information is "fit for use" in organizational tasks and decision making. We also ask you to characterize your company in terms of its quality activities.

Before filling out this questionnaire, you will be told which information to respond about. This set of information was selected because its quality is important to your company.

Your participation in this study is voluntary. If you object to any questions, you may choose not to respond. However, your cooperation is strongly desired and appreciated. This study can only be successful if you carefully and honestly answer the questions.

©Wang, Strong, and Lee, 1996 — 1997

Please do not duplicate or distribute the questionnaire without explicit consent of the authors.

Section 1 Characteristics of the Information
--

For the information you are reporting on in this questionnaire, indicate:

1. The primary type of this information (check one):

- ☐ Financial or Accounting Data
- ☐ Human Resources Data
- ☐ Production or Manufacturing Data
- ☐ Customer, Client, or Patient Data
- ☐ Marketing or Sales Data
- ☐ Clinical Data
- ☐ Other: _____

2. Rate the complexity of the activities for collecting, storing, and using this information.

Very Simple

Very Complex

1 2 3 4 5 6 7 8 9 10

3. Your department (check one):

- ☐ Finance, Accounting
- ☐ Information Systems (MIS)
- ☐ Production, Manufacturing
- ☐ Legal
- ☐ Marketing, Sales
- ☐ Strategic Planning
- ☐ Human Resources
- ☐ Senior Executive
- ☐ Field Operations

4. Your main role relative to this information. Do you primarily (check one):

- ☐ Collect this information
- ☐ Manage those who collect this information
- ☐ Use this information in tasks
- ☐ Manage those who use this information in tasks
- ☐ Work as an information systems
- ☐ Manage information systems professionals

Section 2 Information Quality Assessment

For each statement, indicate the extent to which it is true of this information. "This information" refers to the information or database selected by your company for reporting on in this information quality questionnaire.										
	Not at All					Avg.			Completely	
1. This information is easy to manipulate to meet our needs.	0	1	2	3	4	5	6	7	8	9 10
2. It is easy to interpret what this information means.	0	1	2	3	4	5	6	7	8	9 10
3. This information is consistently presented in the same format.	0	1	2	3	4	5	6	7	8	9 10
4. This information includes all necessary values.	0	1	2	3	4	5	6	7	8	9 10
5. This information is easily retrievable.	0	1	2	3	4	5	6	7	8	9 10
6. This information is formatted compactly.	0	1	2	3	4	5	6	7	8	9 10
7. This information is protected against unauthorized access.	0	1	2	3	4	5	6	7	8	9 10
8. This information is incomplete.	0	1	2	3	4	5	6	7	8	9 10
9. This information is not presented consistently.	0	1	2	3	4	5	6	7	8	9 10
10. This information has a poor reputation for quality.	0	1	2	3	4	5	6	7	8	9 10
11. This information is complete.	0	1	2	3	4	5	6	7	8	9 10
12. This information is presented concisely.	0	1	2	3	4	5	6	7	8	9 10
13. This information is easy to understand.	0	1	2	3	4	5	6	7	8	9 10
14. This information is believable.	0	1	2	3	4	5	6	7	8	9 10
15. This information is easy to aggregate.	0	1	2	3	4	5	6	7	8	9 10
16. This information is of sufficient volume for our needs.	0	1	2	3	4	5	6	7	8	9 10
17. This information is correct.	0	1	2	3	4	5	6	7	8	9 10
18. This information is useful to our work.	0	1	2	3	4	5	6	7	8	9 10
19. This information provides a major benefit to our work.	0	1	2	3	4	5	6	7	8	9 10
20. This information is easily accessible.	0	1	2	3	4	5	6	7	8	9 10
21. This information has a good reputation.	0	1	2	3	4	5	6	7	8	9 10
22. This information is sufficiently current for our work.	0	1	2	3	4	5	6	7	8	9 10
23. This information is difficult to interpret.	0	1	2	3	4	5	6	7	8	9 10
24. This information is not protected with adequate security.	0	1	2	3	4	5	6	7	8	9 10
25. This information is of doubtful credibility.	0	1	2	3	4	5	6	7	8	9 10
26. The amount of information does not match our needs.	0	1	2	3	4	5	6	7	8	9 10
27. This information is difficult to manipulate to meet our needs.	0	1	2	3	4	5	6	7	8	9 10
28. This information is not sufficiently timely.	0	1	2	3	4	5	6	7	8	9 10
29. This information is difficult to aggregate.	0	1	2	3	4	5	6	7	8	9 10
30. The amount of information is not sufficient for our needs.	0	1	2	3	4	5	6	7	8	9 10
31. This information is incorrect.	0	1	2	3	4	5	6	7	8	9 10
32. This information does not add value to our work.	0	1	2	3	4	5	6	7	8	9 10
33. This information was objectively collected.	0	1	2	3	4	5	6	7	8	9 10
34. It is difficult to interpret the coded information.	0	1	2	3	4	5	6	7	8	9 10
35. The meaning of this information is difficult to understand.	0	1	2	3	4	5	6	7	8	9 10

Section 2 (continued) Information Quality Assessment

36. This information is not sufficiently current for our work.	0	1	2	3	4	5	6	7	8	9	10
37. This information is easily interpretable.	0	1	2	3	4	5	6	7	8	9	10
38. The amount of information is neither too much nor too little.	0	1	2	3	4	5	6	7	8	9	10
39. This information is accurate.	0	1	2	3	4	5	6	7	8	9	10
40. Access to this information is sufficiently restricted.	0	1	2	3	4	5	6	7	8	9	10
41. This information is presented consistently.	0	1	2	3	4	5	6	7	8	9	10
42. This information has a reputation for quality.	0	1	2	3	4	5	6	7	8	9	10
43. This information is easy to comprehend.	0	1	2	3	4	5	6	7	8	9	10
44. This information is based on facts.	0	1	2	3	4	5	6	7	8	9	10
45. This information is sufficiently complete for our needs.	0	1	2	3	4	5	6	7	8	9	10
46. This information is trustworthy.	0	1	2	3	4	5	6	7	8	9	10
47. This information is relevant to our work.	0	1	2	3	4	5	6	7	8	9	10
48. Using this information increases the value of our work.	0	1	2	3	4	5	6	7	8	9	10
49. This information is presented in a compact form.	0	1	2	3	4	5	6	7	8	9	10
50. This information is appropriate for our work.	0	1	2	3	4	5	6	7	8	9	10
51. The meaning of this information is easy to understand.	0	1	2	3	4	5	6	7	8	9	10
52. This information is credible.	0	1	2	3	4	5	6	7	8	9	10
53. This information covers the needs of our tasks.	0	1	2	3	4	5	6	7	8	9	10
54. Representation of this information is compact and concise.	0	1	2	3	4	5	6	7	8	9	10
55. This information adds value to our tasks.	0	1	2	3	4	5	6	7	8	9	10
56. The measurement units for this information are clear.	0	1	2	3	4	5	6	7	8	9	10
57. This information is objective.	0	1	2	3	4	5	6	7	8	9	10
58. Information can only be accessed by people should see it.	0	1	2	3	4	5	6	7	8	9	10
59. This information is sufficiently timely.	0	1	2	3	4	5	6	7	8	9	10
60. This information is easy to combine with other information.	0	1	2	3	4	5	6	7	8	9	10
61. This information is represented in a consistent format.	0	1	2	3	4	5	6	7	8	9	10
62. This information is easily obtainable.	0	1	2	3	4	5	6	7	8	9	10
63. This information comes from good sources.	0	1	2	3	4	5	6	7	8	9	10
64. This information is quickly accessible when needed.	0	1	2	3	4	5	6	7	8	9	10
65. This information has sufficient breadth and depth for tasks.	0	1	2	3	4	5	6	7	8	9	10
66. This information presents an impartial view.	0	1	2	3	4	5	6	7	8	9	10
67. This information is applicable to our work.	0	1	2	3	4	5	6	7	8	9	10
68. This information is sufficiently up-to-date for our work.	0	1	2	3	4	5	6	7	8	9	10
69. This information is reliable.	0	1	2	3	4	5	6	7	8	9	10

Section 3: IQ Context Assessment

1...This company has adopted a TQM approach.	N/A	1 2 3 4 5 6 7 8 9 10
2...This company has tools that identify deficiencies with this information.	N/A	1 2 3 4 5 6 7 8 9 10
3...In this company, there are people whose primary job is to assure the quality of information.	N/A	1 2 3 4 5 6 7 8 9 10
4...This company has tools to assure the consistency of this information.	N/A	1 2 3 4 5 6 7 8 9 10
5...In this company, employees view continuous quality improvement as a part of their job.	N/A	1 2 3 4 5 6 7 8 9 10
6...This company uses TQM to control process quality.	N/A	1 2 3 4 5 6 7 8 9 10
7...This company has a specific position or group responsible for information quality.	N/A	1 2 3 4 5 6 7 8 9 10
8...This company solves quality problems using one of the popular quality improvement methods such as one developed by Deming, Juran, or Crosby.	N/A	1 2 3 4 5 6 7 8 9 10
9...In this company, there are designated people whose job is to solve information quality problems.	N/A	1 2 3 4 5 6 7 8 9 10
10...This company has tools to assure the completeness of this information.	N/A	1 2 3 4 5 6 7 8 9 10
11...In this company, there are designated people who are responsible for the quality of information.	N/A	1 2 3 4 5 6 7 8 9 10
12...In this company, employees participate in quality improvement activities.	N/A	1 2 3 4 5 6 7 8 9 10
13...This company has tools to assure the correctness of this information.	N/A	1 2 3 4 5 6 7 8 9 10
14...This company provides software for aggregating, manipulating and summarizing this information.	N/A	1 2 3 4 5 6 7 8 9 10
15...This company is developing a data dictionary to standardize data definitions across different computers or divisions.	N/A	1 2 3 4 5 6 7 8 9 10
16...In this company, employees are able to take actions to improve the quality of information.	N/A	1 2 3 4 5 6 7 8 9 10
17...This company has recently moved this information to a different hardware or software system.	N/A	1 2 3 4 5 6 7 8 9 10
18...In this company, ensuring the quality of this information is the responsibility of those who use the information.	N/A	1 2 3 4 5 6 7 8 9 10
19...This company has new (database) software for managing and storing this information.	N/A	1 2 3 4 5 6 7 8 9 10
20...In this company, it is relatively easy to improve information as needed.	N/A	1 2 3 4 5 6 7 8 9 10

Thank you very much!

References

- [1] Codd, E. F., "A Relational Model of Data for Large Shared Data Banks," *Communications of the ACM*, 13(6), 1970, pp. 377-387.
- [2] Codd, E. F., *The Relational Model for Database Management: Version 2*. Addison-Wesley, Reading, MA, 1990.
- [3] CRG, *Information Quality Assessment Survey: Administrator's Guide*. Cambridge Research Group, Cambridge, MA, 1997.
- [4] CRG, *Integrity Analyzer: A Software Tool for TDQM*. Cambridge Research Group, Cambridge, MA, 1997.
- [5] Lee, Y. W., "Why 'Know Why' Knowledge is Useful for Solving Information Quality Problems." In *Proceedings of Americas Conference on Information Systems*, Phoenix, AZ, 1996, pp. 200-202.
- [6] Lee, Y. W., D. M. Strong, L. Pipino, and R. Y. Wang, *A Methodology-based Software Tool for Data Quality Management* (No. TDQM-97-02). MIT TDQM Research Program, Cambridge, MA, 1997.
- [7] Maier, D., *The Theory of Relational Databases*. 1st ed. Computer Science Press, Rockville, MD, 1983.
- [8] Pipino, L., Y. W. Lee, and R. Y. Wang, *Measuring Information Quality* (No. TDQM-97-04). MIT Sloan School of Management, Cambridge, MA, 1998.
- [9] Wang, R. Y., "A Product Perspective on Total Data Quality Management," *Communications of the ACM*, 41(2), 1998, pp. 58-65.
- [10] Wang, R. Y., Y. L. Lee, L. Pipino, and D. M. Strong, "Manage Your Information as a Product," *Sloan Management Review*, 39(4), 1998, pp. 95-105.
- [11] Wang, R. Y., and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems (JMIS)*, 12(4), 1996, pp. 5-34.