An MIS Data Quality Methodology Based on Optimal Error Detection

David B. Paradice and William L. Fuerst

SYNOPSIS

As organizational information systems increase in use, cost, complexity, and importance, the quality of the data upon which decisions are based becomes critical. Presently, quantitative measures of an organization's data quality do not exist, with organizations relying on periodic audits of selected applications to evaluate the accuracy of their systems. In this paper, a quantitative measure is developed by formulating the error rate of stored MIS records, whereby MIS records are classified as being either "correct" or "erroneous." This approach has several desirable characteristics including: (1) specification of a benchmark for comparing the actual error rate with the theoretically smallest attainable error rate, and (2) specification of a method for allocating records to the categories in an optimal manner.

Following development of the formulation, a methodology is provided for assessing an organization's data quality. Since quality control methodologies are beneficial only when appropriately applied in the specific organizational setting at hand, guidelines are given for deriving values for the parameters required in the formulation. These values can be obtained from a variety of sources, both mechanized and manual. The paper concludes with a discussion of the relationship between the parameters in the formulation and the verification mechanisms designed to insure data quality.

Key Words: Data quality, Accounting controls, Error measurement, Error classification mechanisms

THE importance of data quality in management information systems (MIS) increases daily. As organizational information systems increase in use, cost, and complexity, the quality of the data upon which decisions are based becomes critical. Erroneous data may lead to decisions and actions that have severe consequences. The integrity of customer credit ratings, manufacturing processes, patient care, investment portfolios, and employee compensation can be jeopardized by faulty data. For example, research by Laudon (1986) involving one sample of criminal information system records indicated that more than 14,000 persons were at risk of being inappropriately detained and perhaps arrested every day due to inaccurate data.

Accounting professionals have been concerned with data quality measurement for some time. Yu and Neter (1973) developed a statistical approach to the measurement of errors in outputs of internal control systems. Cushing (1974) developed a mathematical model of the accounting internal control system and measures of reliability and cost. Johnson et al. (1981) and Groomer and Murthy (1989) addressed data quality as it relates to audit populations. Hamlen (1980), Stratton (1981), and Fields et al. (1986) presented models of the internal control process which responded to guidelines and regulations calling for auditors to evaluate management's effort to assure that accounting data was correct. Ballou and Pazer extended the models developed in the accounting literature (1985a), analyzed cost/

David B. Paradice and William L. Fuerst are at the College of Business Administration and Graduate School of Business at Texas A&M University.

The authors would like to thank the editor and the four anonymous reviewers for their helpful comments during the review process.

per 1983, 1985; Emery 1969), wherein data are raw materials and information (i.e., processed data) are outputs, the corresponding notions of quality control have not surfaced in information systems methodologies. This situation may have occurred because data, unlike most raw materials, is not consumed when processed and therefore may be reused repeatedly. While the cost of data "waste" may be zero, the cost of using inaccurate data certainly may be large.

Records in an organization's information system may contain hundreds of data items each. As the number of data items per record increases, determining errors in these data items or erroneous combinations of data item values becomes quite difficult. In addition, audits of specific applications are performed infrequently due to the cost involved. Error detection is performed most commonly by users on an ad hoc basis as they observe peculiarities in data item values. Unfortunately, humans have a very low threshold for handling complexity (Volonino and Kirs 1988), frequently exhibiting a range of cognitive biases which preclude either efficient or effective performance. Any methodology for assessing data quality must consider the inherent complexity in typical MIS records.

Our concern in this paper focuses on minimizing the introduction of erroneous records into an MIS. In particular, we focus on combinations of data values which, when considered in total, make a record atypical. Consequently, this formulation does not focus on "erroneous" transactions which are routinely handled by introducing "compensatory" transactions (although the formulation certainly accommodates this case). The formulation we present below does not provide guidance on how to deal with erroneous records already in the system. However, if corporations can avoid storing erroneous records, they will have taken a good "first step" toward reducing the amount of incorrect data in the MIS. In the remainder of the paper, we first present a methodology for MIS data quality control by formulating the error rate of records stored in an MIS as a statistical classification problem. Following the methodological development, we provide an algorithm for applying the methodology.

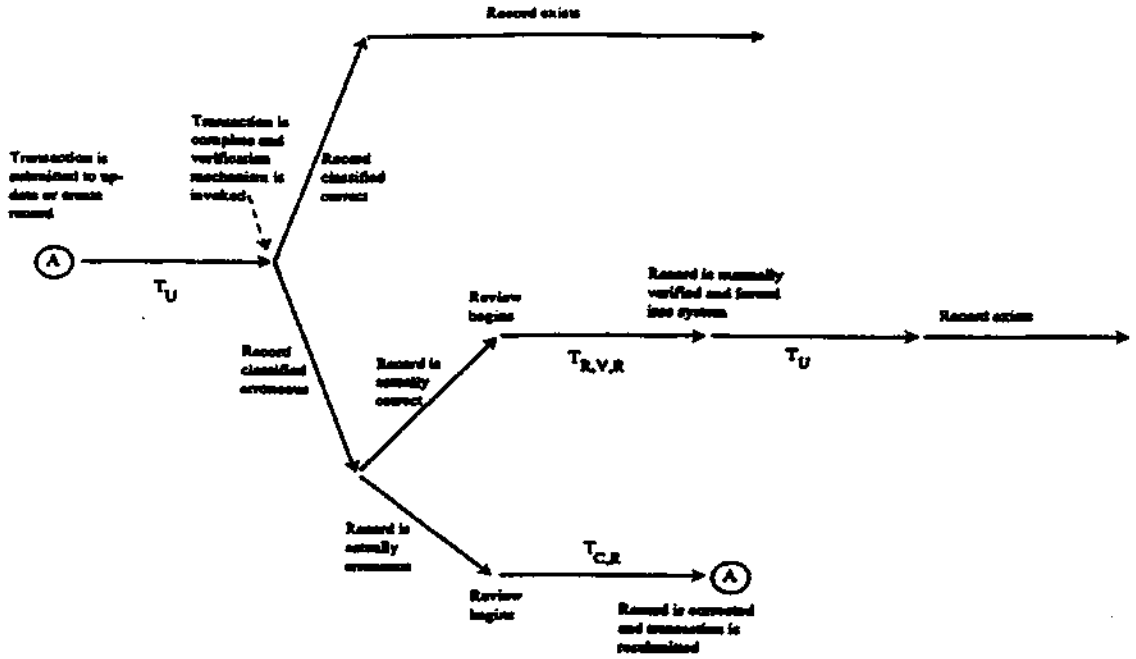
Before proceeding with the formulation development, definitions of several terms are needed for clarity. An **MIS record** is the collection of data items that is stored for the long term on a medium like disk. **Transactions** are the result of accounting events which require the processing of stored MIS records. For the purposes of this paper, transactions will not be referred to as records in order to provide clarity in differentiating between stored MIS records and relatively transient transactions. An MIS record is **updated** only after it is written back to the disk after the transaction is applied to it. Although some processing of the record occurs in the computer's main memory, the record is not considered updated until it is changed on disk. **Verification** refers to a procedure which evaluates the "correctness" of the record's data after the transaction has been applied but before the record is written back to disk. The next section further explores the application of transactions to MIS records and the updating of those records after verifying their contents.

STORED MIS ERROR RATES

Calculating the error rate of stored MIS records requires determining the probability that a randomly chosen record in an MIS is correct (i.e., that it contains data which is accurate). The interval of time (T) that a record exists in an MIS may be partitioned into two periods: a period E when the record is erroneous and a period C when the record is correct. The partition may occur anywhere in the interval. As pointed out by Morey (1982), E/T equals the percentage of time a randomly chosen record is in error. The goal is to determine ways to minimize the expected value of this quantity (E/T).

Due to updates, MIS records may have many lives as defined by the partitioning

FIGURE 1
TIME DIMENSION OF UPDATE AND VERIFICATION PROCESSES



credit terms in a customer record provide better discriminators than customer name and address. However, in certain circumstances, demographic data might be used as a discriminator because of the type of transaction being processed. For instance, in a hospital information system, a very good discriminator for the correctness of a record reflecting an ordered pregnancy test would be the demographic variable SEX. If this variable contains the value "male" the record is likely to be incorrect. Thus, this example illustrates that the verification mechanism may actually consist of a set of evaluations tailored for different types of transactions which may occur. In this case, any one of the evaluations may be sufficient to cause the verification mechanism to flag the modified record as erroneous. We use $P_{CE}(x)$ to denote the probability that a record will be classified as "erroneous" (i.e., the verification mechanism calculates a value in the range R_E), and $P_{CC}(x)$ denotes the probability that a record will be classified as "correct" (in the range R_C). Also, P_{AE} represents the probability that a record is actually erroneous, and P_{AC} represents the probability that it is actually correct.

Figure 2a presents with hypothetical values the nature of the problem in the univariate case. Ideally, the fields that compose x will be chosen so that the two distributions do not overlap. In this case, there can be no **misclassification**. The figure shows the more likely case, though, where the two distributions overlap somewhat. We are interested in the area labeled "Region of Misclassification," for this is the area in which the verification mechanism determines correct records are more similar to erroneous ones, and vice versa. (The position of the line dividing R_E and R_C is discussed shortly.)

Figure 2b illustrates how the choice of fields to be included in x affect the notion of a correct or erroneous record. In this

simple case, two attributes have been chosen to compose x. The shaded area represents typical combinations of values for these attributes. At point A, both attributes have values which are typical. A record with these values would be classified as correct. Assume a transaction modifies the value of attribute 1, thus moving the combination of values to point B. Still, the combination is typical, and such an update would result in a "correct" record. However, a subsequent update to the same record by modifying attribute 2 (or an update to the original record by modifying both attributes) such that point C represents the modified record would result in the record being classified as "erroneous." The combination of values is no longer "typical."

Using these terms, we can then formulate the stored MIS error rate (e_{MIS}) as follows (Morey 1982):

$$e_{MIS} = E\left[\frac{\min(T_U, T)}{T}\right] P_{AE} \int_{R_C} P_{CE}(x)\,dx \tag{1a}$$

$$+ E\left[\frac{\min(T_U + T_{C.R}, T)}{T}\right] P_{AE}\left[1 - \int_{R_C} P_{CE}(x)\,dx\right] \tag{1b}$$

$$+ E\left[\frac{\min(T_U, T)}{T}\right] P_{AC}\left[1 - \int_{R_E} P_{CC}(x)\,dx\right] \tag{1c}$$

$$+ E\left[\frac{\min(T_U + T_{R.V.R}, T)}{T}\right] P_{AC} \int_{R_E} P_{CC}(x)\,dx \tag{1d}$$

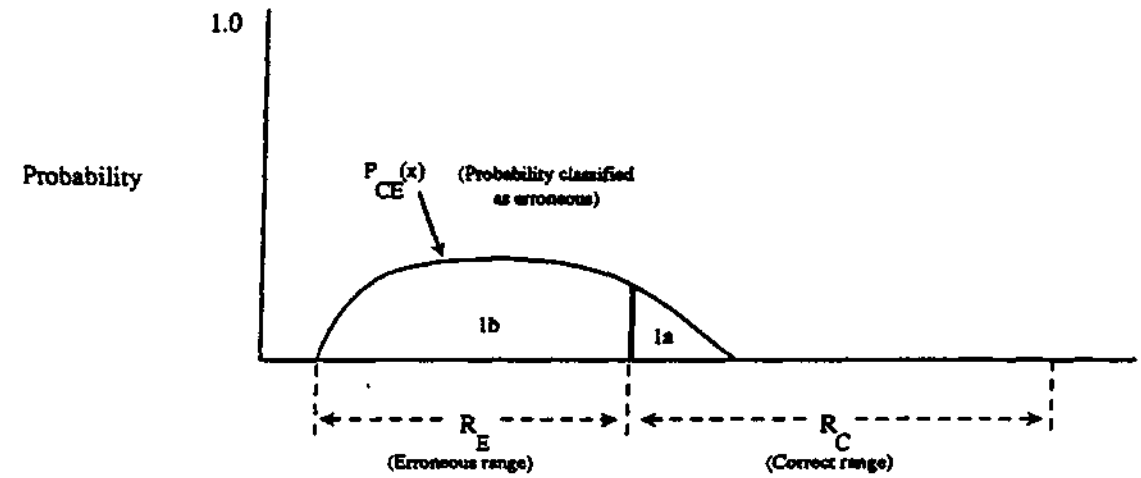
where min (x, y) denotes the appropriate value to use is the minimum of x and y.

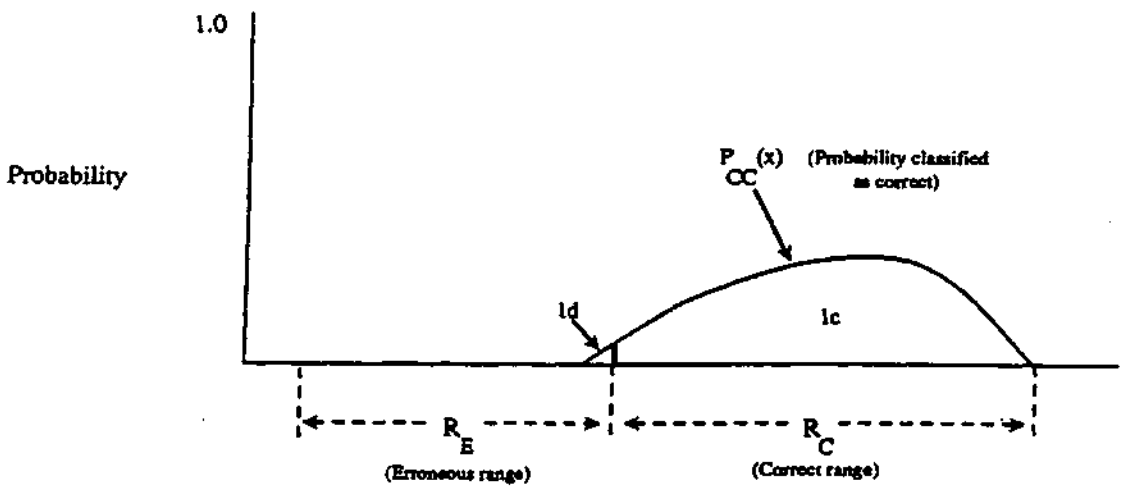
At first glance, the formulation may appear to be incorrect, as the subscripts are somewhat mixed throughout the formula. However, we are interested in the *mis*classification that occurs. Therefore, we focus on areas where correct records are considered similar to erroneous ones, and vice versa. The subscripts reflect this focus.

Figure 3 is a decomposition of figure 2a and helps illustrate the formulation. The term (1a) accounts for erroneous records which are misclassified as correct. We are interested in the probability that the record is classified as erroneous ($P_{CE}(x)$) when the records may have the characteristics of being correct (R_C). Thus, the integration occurs over this area. Because we are formulating an error rate, the term is "weighted" by the probability that the record is actually in error (P_{AE}) and by the time spent processing it. The terms containing time values may be further explained as follows. In most cases, the Ts with subscripts will be less than T (i.e., update, review, correction, and resubmission processes are shorter in duration than the total life of the record being processed). Since we are formulating an error rate, and since we are interested in choosing a record at random, we weight the likelihood of selecting a record by the percentage of time it is processed. In cases where processing time is equal to or exceeds the life of the record, these terms reduce to T/T = 1, maintaining the applicability of the formulation (i.e., ensuring the formulation "behaves" in extreme cases).

Term (1b) accounts for the remaining records likely to be classified erroneous. These are the erroneous records properly rejected by the system. This term is also weighted by the probability that the record is actually in error and by a slightly different processing time. The processing time in (1b) accounts for the additional time needed to correct and resubmit the record. Terms (1c) and (1d) may be explained in a similar manner. Term (1d) accounts for correct records which are misclassified as er-

FIGURE 3
FOUR POTENTIAL CONDITIONS OF AN MIS RECORD





roneous. This term is weighted by the probability that the record is actually correct and a processing time that reflects the additional time spent to review, verify, and resubmit a record inappropriately rejected by the verification mechanism. Term (1c) accounts for the remaining records likely to be classified as correct. These are the correct records properly accepted by the system.

Through algebraic manipulation and recognizing that the expectation terms reduce to constants (see appendix), Morey's formulation may be rewritten as:

$$e_{MIS} = A\ P_{AE} \int_{R_C} P_{CE}(x)\,dx + B\ P_{AC} \int_{R_E} P_{CC}(x)\,dx$$

$$+ P_{AE}\left[\frac{\min(T_U + T_{C,R}, T)}{T}\right] + P_{AC}\left[\frac{\min(T_U, T)}{T}\right] \quad (2)$$

In classical statistical classification theory, the quantities represented by A and B above often represent misclassification costs (Johnson and Wichern 1988). In our formulation, these quantities represent the percentages of time that a record is classified as "erroneous" (R_E) or "correct" (R_C). Thus, the stored MIS error rate may be defined as the percentage of time of misclassification plus additional processing time to address the misclassification. These times are weighted by the probability that the record is actually erroneous (P_{AE}) or actually correct (P_{AC}), as follows:

$$P_{AE}\left[\frac{\min(T_U + T_{C,R}, T)}{T}\right] + P_{AC}\left[\frac{\min(T_U, T)}{T}\right] \quad (3)$$

The role of these terms may be explained best as follows. In the case of perfect classification, the expected percentage of time of misclassification is equal to zero. That is, if all records with errors were classified as erroneous, and all records with no errors were classified as correct, the expected percentage of time of misclassification would be zero. In this case of perfect classification, the stored error rate (e_{MIS}) is equal to only the time necessary to correct and resubmit a transaction that would have introduced an erroneous record into the MIS (the first term in (3)) and the time required to process transactions that result in correct modified records (the second term in (3)). In the absence of perfect classification, additional processing time would be incurred.

Significantly, (3) can also serve as a diagnostic device, for it represents the smallest error rate that can possibly be attained, assuming that errors will always be introduced into the system, and that selection of a subset of fields from a record is unlikely to result in perfect classification. In other words, if an MIS installation's error detection procedures produce an error rate "close" to the rate specified in (3), the installation is operating as well as can be expected given the effectiveness of the fields (x) used to implement the verification mechanism. If the error rate is not close, then the verification mechanism can be adjusted by determining additional data items (or a new combination of data items) to include in x for verification to improve the classification. (We provide an algorithm for constructing this mechanism below.) This benchmark gives MIS management, for the first time, a clear measure for evaluating current verification procedures and proposed changes.

The parameters of the formulation define a threshold value for classifying transactions, i.e., for determining the boundary between R_E and R_C. To minimize the stored error rate, the verification mechanism should classify a record as erroneous or correct based on the following decision rule:[1]

$$\text{erroneous if } \frac{P_{CE}(x)}{P_{CC}(x)} \geq \frac{B\ P_{AC}}{A\ P_{AE}} \quad \text{and}$$

$$\text{correct if } \frac{P_{CE}(x)}{P_{CC}(x)} < \frac{B\ P_{AC}}{A\ P_{AE}} \quad (4)$$

These "decision rules" for classification yield an intuitive application. Consider the case of a simple inadmissibility verification

[1]The basis for this decision rule follows from the nonnegative nature of the components of the formulation and the fact that the misclassification only depends on x. Thus, the misclassification is minimized when R_E contains the points for which $P_{AE}P_{CE}(x)$ exceed $P_{AC}P_{CC}(x)$. Algebraic manipulation leads to the form above. For a detailed proof, see Johnson and Wichern (1988, 485).

where a value must not exceed an upper bound, like a sales application where customer balance must not exceed credit limit. In this case, a record containing an inadmissible value, i.e., the exceeded credit limit, is not accepted. Therefore, the probability that a valid record has a value outside of the proper range ($P_{CC}(x)$) is zero. If division by zero is defined as resulting in an infinitely large value, then the record is classified as erroneous using the decision rule in (4). Thus, simple admissibility verification classifies some records according to the optimal rule. However, the optimal formulation implies that possibly many more records should be classified as "erroneous" if one wishes to minimize the MIS stored error rate.

Quantities representing costs may be added as an extension to this formulation. We assume there are no interesting costs associated with the cases wherein correct records are appropriately accepted and incorrect records are appropriately rejected by the verification mechanism. The system is functioning properly in these cases. The cases where misclassification occurs, however, are characterized by potentially devastating costs.

One case of misclassification occurs when incorrect data is accepted as correct. There are several examples which illustrate this type of misclassification. Consider the manufacturing company which processes products that are out of specification due to an error when modifying its bill of materials file. Or the distributing company that miscalculates discount amounts because of errors in customer records concerning payment terms. Both of these examples, and many like them, can be very costly.

The other case of misclassification occurs when correct data is rejected as incorrect, and it can also have costly results. Because of the additional time required to review and expedite this type of misclassification, the most drastic kinds of examples involve time-sensitive situations. That is, by not being able to respond quickly to some situation, costs are incurred. For example, consider the investment company whose purchase of stock is delayed due to this type of misclassification, only to find out later that the price of the stock has significantly increased during the time to correct the inappropriate classification. Or the retailer whose processing of payables was delayed because records were inaccurately classified as erroneous, causing discount deadlines to be missed. With these types of misclassifications, the costs include those associated with (1) an expert's time to "validate" the already correct data, (2) the effort to enter the transaction a second time, and (3) any opportunity that is delayed or missed due to the data's unavailability.

We can incorporate cost terms into the formulation, where K_A represents the cost of accepting erroneous records and K_R represents the cost of rejecting correct records, as follows:

$$e_{MIS} = K_A A P_{AE} \int_{R_C} P_{CE}(x)\,dx + K_R B P_{AC} \int_{R_E} P_{CC}(x)\,dx + P_{AE}\left[\frac{\min(T_U + T_{C,R}, T)}{T}\right] + P_{AC}\left[\frac{\min(T_U + T)}{T}\right] \quad (5)$$

In this form, the optimal decision rule becomes:

$$\text{erroneous if } \frac{P_{CE}(x)}{P_{CC}(x)} \geq \frac{B K_R P_{AC}}{A K_A P_{AE}}$$

$$\text{correct if } \frac{P_{CE}(x)}{P_{CC}(x)} < \frac{B K_R P_{AC}}{A K_A P_{AE}} \quad (6)$$

The costs influence the optimal decision rule as expected. When the cost of using erroneous data is greater than the cost of examining the record a second time, the decision rule becomes more sensitive to erroneous records and rejects more frequently.

When the cost of (re)examining records exceeds the cost of using erroneous data, records are not rejected as often according to the optimal rule.

APPLICATION OF THE METHODOLOGY

The above formulation describes the methodology for assessing data quality in an information system. This methodology provides a different type of measure of data quality than traditional auditing procedures. The quantified measure of the stored MIS error rate obtained from the methodology is based on many variables as presented above, including a variety of time dimensions and experts' opinions. In order to apply this methodology, values for the variables in the formulation must be obtained. In most cases, the values of details like record frequency, processing time, and record life span can be obtained fairly easily from computer operations or database management system records. In other cases, computer specialists or application experts will be required to provide estimates. These estimates will often require the use of some type of statistical sampling of the data records. A common technique for such sampling (Gallegos et al. 1987) and one that is particularly well-suited for this methodology is random attribute sampling. With this sampling technique, experts will be able to specify attributes of the MIS records to be selected, allowing the experts to focus on fields within the records which are critical from a data quality standpoint.

The following discussion presents each of the factors considered by the methodology and a proposed method for arriving at the values.

Estimate the prior probability that a transaction results in a record that is in error (P_{AE}) and the prior probability that a transaction results in a record that is correct (P_{AC}).

The intrinsic error rate (P_{AE}) is the rate at which erroneous records are created due to flaws inherent in the system generating the transactions. An estimate of the intrinsic error rate may be obtained by evaluating a sample of transactions and determining the number that produce incorrect records. Depending on the cost of this evaluation process, one might evaluate 100 or 1000 transactions, but certainly a better intrinsic error rate estimate will be obtained using a larger sample. Once the intrinsic error rate (P_{AE}) is estimated, the probability that a transaction produces a correct record (P_{AC}) is $1 - P_{AE}$.

Determine the time that a record typically exists in the system (T).

The amount of time that a record exists in a system will vary according to the applications supported by the system. This time may be measured (1) in minutes, as for a record tracking a specific flight through the air traffic control system; (2) in days, as for a record tracking a specific person through a hospital administration system; or (3) in years, as for a record tracking a specific person through an organizational personnel system. These times can be obtained by referring to time stamps on records in the specific application.

Obtain the minimum elapsed time from submission of a transaction to the system until it updates a record (T_U).

In most interactive data processing environments, the minimum elapsed time from submission of a transaction to the system until it updates a record is negligible. This time may be measured in milliseconds, and can be obtained by accessing the system clock immediately before the transaction is submitted and immediately after the modified record is stored. Almost all database management software currently available provide mechanisms for obtaining this information (Soderlund 1986).

In a batch environment, however, this time could conceivably be measured in much longer intervals (e.g., hours). This time can also be obtained as above, since the transaction must still be created and ultimately submitted to the system.

Estimate the additional processing time delay to review, correct and resubmit transactions which modified records resulting in erroneous data, and properly failed verification ($T_{C,R}$).

This additional processing time is dependent on the reasons that the modified record failed the verification mechanism. In the simplest case, a modified record fails because a single value is unacceptable (i.e., an inadmissible value). In the worst case, a complex combination of values which are each individually acceptable but in combination are unacceptable will cause the modified record to fail the verification process. These cases must be corrected by an expert(s) familiar with the data, a process that could take minutes to days depending on the availability of the expert and the nature of the error. Still, an estimate of the additional time can be obtained by monitoring over time how long transactions are out of the system in order to be corrected.

Estimate the additional processing time to manually review, validate, and resubmit to the system any intrinsically correct transactions which modified records resulting in correct data but failed the verification process ($T_{R,V,R}$).

As above, this additional processing time could vary depending on the reason for failure. In the simplest case, a transaction is known to be atypical, yet correct, from its origin. For example, a transaction to update a record in a customer information system could cause the customer's credit limit to be exceeded. The system administrator knows *a priori* that this transaction is going to cause the modified record to fail verification, but he also knows that this transaction is correct (perhaps due to a new policy that allows selected customers with excellent credit records to make a large purchase on a one-time basis). The additional processing time in this case is negligible, and the record is forced into the system.

However, the worst case occurs when a complex combination of data values are applied in a transaction. An expert opinion must be sought to determine the validity of the transaction. This process may require extensive review of source data, transcription devices, and measuring instruments (human and mechanical). Minutes, hours, or days may be required to determine the validity of the transaction. As above, an estimate of the additional time can be obtained by monitoring over time how long transactions are out of the system in order to be validated.

Estimate the probability density function for erroneous records ($P_{CE}(x)$) and the probability density function for correct records, ($P_{CC}(x)$), based on a subset of the data items from the record.

Selecting the best candidate fields to be included in the verification mechanism could be determined by having experts review sets of records. The experts would classify the records as correct or incorrect, based on their expertise and the degree to which the data is consistent with their expectations. This reasoning process would emphasize the values used to make the classification. The results of the experts' classifications would be compared with the actual status of the records to estimate these distributions.

These distributions form the conceptual basis for the verification mechanism. As noted earlier, these fields should be chosen carefully. Fields containing data such as name and address are rarely appropriate. Fields identified by experts as being most useful in "predicting" a record's "correctness" are the "discriminating" fields from the record.

Theoretically, these estimates require determining the likelihood of all combinations of possible values for some specified number of fields. However, selective use of general categories could make estimating these distributions easier. For example, in a payroll system, instead of determining the likelihood of a combination of annuity plans for each individual, one would determine the likelihood of the plans for categories of individuals such as "executive management" and "divisional management." Or in-

stead of determining the likelihood of all possible payment amounts for each vendor in a purchasing system, one would determine the likelihood of payment ranges for "classes of vendors."

Measure the actual error rate of the system over time, and compare this error rate to the value defined in (3).

The actual error rate over time is determined by monitoring the system and simply counting the number of erroneous records that are found. The average of the periodic measurements provides an estimate of the actual error rate over time.

Once implemented, if the actual error rate exceeds the value defined in (3), re-examine the definition of $P_{CE}(x)$ and $P_{CC}(x)$.

Formulation of the "remainder" assumes perfect classification. Thus, improvement in the actual classification method (i.e., the verification mechanism) should force the actual error rate to approach the value of the "remainder." Conceptually, improving the classification method requires adjusting $P_{CE}(x)$ and $P_{CC}(x)$ so that these probability density functions become more distinct (i.e., less overlapping). More pragmatically, the verification mechanism may be improved by adding or deleting fields from the mechanism that are poor discriminators or by tightening the constraints on some field values given the values of other fields. As in any quality control process, the assumptions made and the parameters employed should be evaluated periodically for appropriateness. In this case, introduction of new procedures, new systems, or new data collection techniques could affect the actual probabilities. Reviewing and re-estimating the parameters of the formulation periodically would be prudent, and such action is especially worthwhile if the stored MIS error rate should jump suddenly.

SENSITIVITY ANALYSIS OF THE PARAMETERS IN THE MECHANISM

To evaluate the general impact of the parameters on verification mechanisms, we conducted a series of simulation scenarios. In these scenarios, threshold values for the optimal decision rule in (4) were calculated examining a range of values for the parameters in the formulation. The following parameters were manipulated: the probability that a transaction modification results in a record that is in error (P_{AE}); the length of time the record is in the system (T); and the additional processing time required due to classifying records as erroneous ($T_{C,R}$ and $T_{R,V,R}$). These simulations were performed for two reasons: (1) to illustrate the relationship between the parameters in the formulation and verification mechanisms, and (2) to determine if the formulation was reasonable from a practical standpoint, confirming intuitively appealing verification practices.

Values for the parameters in the formulation were selected to represent typical business applications. Three hundred twenty-four different scenarios were simulated based on two different sets of calculations, as depicted in Figure 4. Values for T_U, $T_{C,R}$, and $T_{R,V,R}$ were selected in Set #1 to represent a batch environment, with the Set #2 being more representative of an online environment. The range of values for T in both sets were selected to represent typical accounting cycles, where a record might exist in the data base for a week, one-half month, one month, and so on. Both sets also used the same values for P_{AE}, which were selected to represent potential error rates ranging from .05 to .20. For each of the 324 scenarios, the optimal decision rule was calculated for all possible combinations of parameter values provided.

In each of the two sets of calculations, the values for T_U, $T_{C,R}$, and $T_{R,V,R}$ closely approximated each other, using the same unit of measure. That is, if the time (T_U) to process a correct transaction is one day, the time ($T_{C,R}$) to rectify an error for such a transaction approximates that processing time. As shown in figure 4, the processing time (T_U) was held constant at one day in the first set of calculations, and the time ($T_{C,R}$) to rectify an error was one, three, and

FIGURE 4
PARAMETER VALUES USED TO CALCULATE THRESHOLD VALUES

Set #1

T_U	= 1 day
$T_{C,R}$	= 1, 3, 7 days
$T_{R,V,R}$	= 1, 3, 7 days
T	= 7, 15, 30, 90, 365, 3650 days
P_{AE}	= .05, .10, .20

Set #2

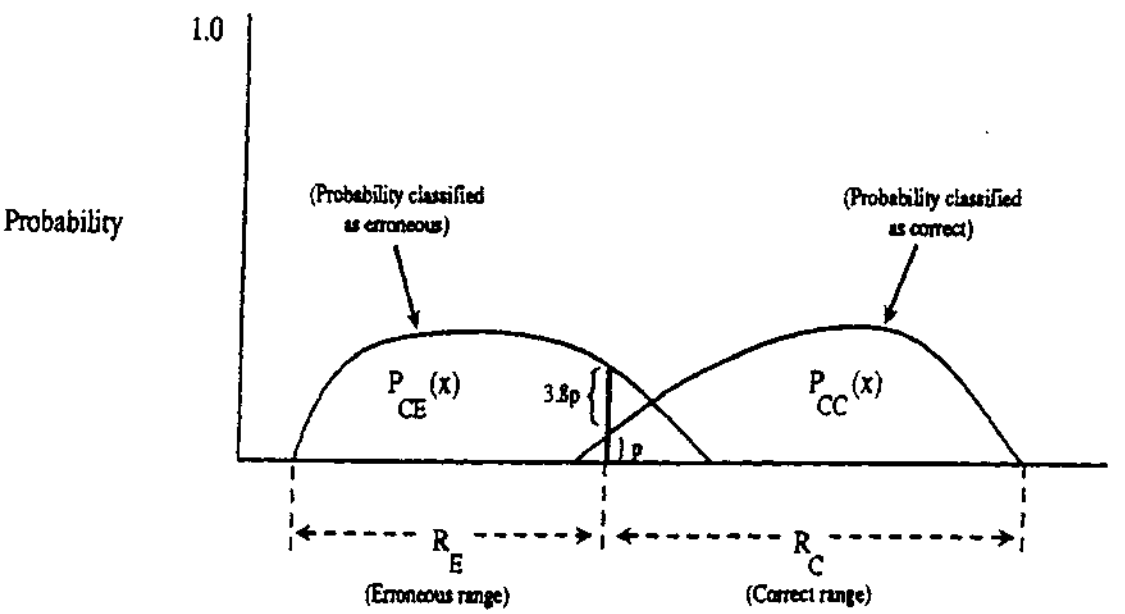
T_U	= 60 mins
$T_{C,R}$	= 60, 120, 180 mins
$T_{R,V,R}$	= 60, 120, 180 mins
T	= 7, 15, 30, 90, 365, 3650 days
P_{AE}	= .05, .10, .20

seven days. The time ($T_{R,V,R}$) to re-enter intrinsically correct transactions that modified records that failed verification was also one, three, and seven days. The time (T) the record remained in the system was 7, 15, 30, 60, 90, 365, 3650 days in Set #1. The unit of measure for the values for T_U, $T_{C,R}$, and $T_{R,V,R}$ was changed from days to minutes for Set #2. The values for the probability that a transaction modification resulted in a record that was in error for all sets of calculations was .05, .10, and .20. For example, taking T_U equal to one day, $T_{C,R}$ equal to one day, $T_{R,V,R}$ equal to one day, T equal to seven days, and P_{AE} = .05, the threshold value is 3.8. This value indicates the probability of the modified record being erroneous must be at least 3.8 times greater than the probability of it being correct for the modified record to be classified erroneous according to (4).

Figure 5 illustrates the placement of the partition between R_E and R_C given these parameters and the calculation in (4). The partition is located where the value of the evaluation of $P_{CE}(x)$ is 3.8 times larger than the evaluation of $P_{CC}(x)$. Letting P_{AE} = .10, and holding the other parameters constant, the threshold value drops to 1.8. In this case, the partition moves to the right and R_E expands as expected.

The threshold values of the optimal decision rule in (4) were calculated as illustrated above for all of the combinations of

FIGURE 5
PLACEMENT OF PARTITION BETWEEN ERRONEOUS AND CORRECT RANGES GIVEN PARAMETERS



Values Generated by Verification Mechanism

values presented in figure 3. These values were then plotted to show the relationship between the threshold values and the methodology parameters (P_{AE}, T, $T_{C,R}$, and $T_{R,V,R}$). For each parameter, a pattern was noted, as illustrated in Figure 6. Each pattern, represented by the shape of the curve, is a function of the combination of parameters used in calculating the threshold values.

As can be seen, each pattern is intuitively appealing, which supports the practicality of our formulation. In Figure 6, the top pattern indicates that as the intrinsic error rate increases, the threshold value decreases, meaning that the optimal decision rule classifies more records as erroneous. Thus, the intuitively appealing consequence of this pattern is that the verification mechanism rejects more records as more erroneous records enter the system.

The middle pattern in Figure 6 illustrates the impact of record life spans. As the record remains in the system for longer periods of time, the threshold value decreases, indicating that the optimal decision rule classifies more records as erroneous. This is an intuitive result. If records are incorrect, then over time the stored error rate will increase as those records accumulate due to their longer life span. Consequently, the verification mechanism based on discriminating data becomes critical as the record life increases.

The lower pattern in Figure 6 illustrates the effect of additional processing times. As the additional processing times increase, the threshold value also increases, meaning that the optimal decision rule is less likely to classify a record as erroneous. From a practical standpoint, the verification mechanism must be "tempered" by the time required for additional processing, either to correct transactions that modified a record resulting in erroneous data or to resubmit intrinsically correct transactions which modified a record resulting in correct data but that failed verification. If the verification mechanism is too stringent and the additional processing time is too long, one runs the risk of having records never entering the system (before their useful life expires).

CONCLUSIONS

Data quality is a fundamental concern in today's organizations. As the reliance on computer-based information systems grows, the need for maintaining accurate records grows correspondingly. Yet managers of information resources are often left without quantitative measures of the quality of the data in their information systems.

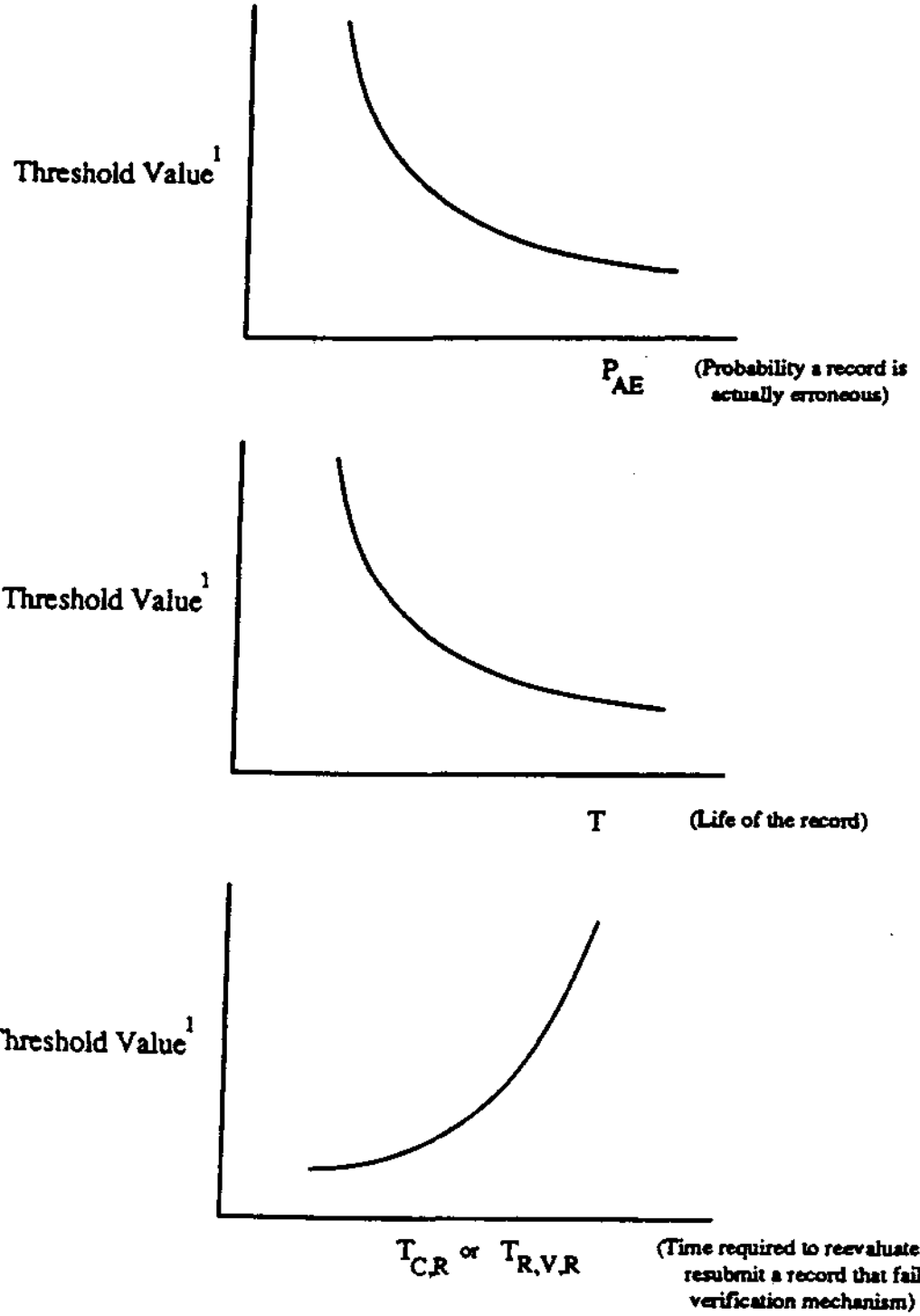
This paper presented a methodology for measuring an information system's error rate as a means of evaluating an organization's data quality. This quality control methodology was formulated in terms of a statistical classification problem, where a verification mechanism classifies records into two categories: correct and erroneous.

After formulating the methodology, the practical implications of applying it become important. We have presented a rather extensive discussion on obtaining values for the various parameters in the formulation. These values can be obtained either from processing details kept in computer operations and database management system records, or from computer specialists and applications experts. These values, when applied to the methodology and compared with the organization's error detection procedures, provide a measure to determine if an organization is operating near optimal conditions.

Also presented in this paper are the relationships among the parameters in the formulation and the verification mechanisms. Through a series of simulated scenarios using ranges of data for the various parameters, three relationships were shown to exist. First, as the probability that records will be in error increases, the optimal decision rule classifies more records as erroneous. Second, as the time that records will be in the system increases, erroneous classifications increase. In both cases, the desire to classify records as erroneous results in more rigorous verification.

However, the third relationship observed argues for less rigorous verification.

FIGURE 6
TRADE-OFFS OF CLASSIFICATION PROBABILITY TO METHODOLOGY PARAMETERS



[1]Threshold value is ratio of probability of classifying a record as erroneous to the probability of classifying it as correct.

As the time increases for additional processing to correct errors or to determine why correct records were classified as erroneous, the optimal decision rule classifies fewer records as erroneous. The effect of this relationship is a dampening of the previous two.

Significantly, all of these relationships, as calculated in the formulation, result in intuitively appealing conclusions regarding the use of verification mechanisms. Without this appeal, the basis of the formulation could be questioned from a practical standpoint.

The research may be extended in several ways. Research is needed to determine how verification methods may vary under different conditions. Distributions of "erroneous" fields may vary given the type of data stored. For example, fields in financial records may tend to exhibit one error distribution, while fields in inventory records exhibit another. Additionally, research needs to be oriented toward qualitative data, since much of the data used by managers is qualitative. Analyses of the effectiveness of control devices (as established in audits) such as limit tests, redundant data checks, range checks, and compatibility checks are also needed. Finally, research is needed to compare and contrast various strategies and the conditions affecting those strategies. Investigations such as these will lead to a better understanding of ways to maintain MIS data quality.

APPENDIX

We begin with

$$e_{MIS} = E\left[\frac{\min(T_U, T)}{T}\right] P_{AE} \int_{R_C} P_{CE}(x)\,dx \tag{A1a}$$

$$+ E\left[\frac{\min(T_U + T_{C.R}, T)}{T}\right] P_{AE}\left[1 - \int_{R_C} P_{CE}(x)\,dx\right] \tag{A1b}$$

$$+ E\left[\frac{\min(T_U, T)}{T}\right] P_{AC}\left[1 - \int_{R_E} P_{CC}(x)\,dx\right] \tag{A1c}$$

$$+ E\left[\frac{\min(T_U + T_{R.V.R}, T)}{T}\right] P_{AC} \int_{R_E} P_{CC}(x)\,dx \tag{A1d}$$

$$- \left\{E\left[\frac{\min(T_U, T)}{T}\right] - E\left[\frac{\min(T_U + T_{C.R}, T)}{T}\right]\right\} P_{AE} \int_{R_C} P_{CE}(x)\,dx$$

$$+ \left\{E\left[\frac{\min(T_U + T_{R.V.R}, T)}{T}\right] - E\left[\frac{\min(T_U, T)}{T}\right]\right\} P_{AC} \int_{R_E} P_{CC}(x)\,dx$$

$$+ P_{AE} E\left[\frac{\min(T_U + T_{C.R}, T)}{T}\right] + P_{AC} E\left[\frac{\min(T_U, T)}{T}\right] \tag{A2}$$

To simplify this expression, we can substitute the terms A and B for the bracketed terms in (A2), as follows:

$$e_{MIS} = A\,P_{AE} \int_{R_C} P_{CE}(x)\,dx + B\,P_{AC} \int_{R_E} P_{CC}(x)\,dx$$

$$+ P_{AE} E\left[\frac{\min(T_U + T_{C.R}, T)}{T}\right] + P_{AC} E\left[\frac{\min(T_U, T)}{T}\right] \tag{A3}$$

In classical statistical classification theory, the quantities represented by A and B above often represent misclassification costs. For instance, Johnson and Wichern (1988) define the expected cost of misclassification (ECM) as:

$$ECM = C(2 \mid 1)\,P_1 \int_{R_2} f_1(x)\,dx + C(1 \mid 2)\,P_2 \int_{R_1} f_2(x)\,dx \tag{A4}$$

This formula can be translated into our formulation by substituting

A for C(2 | 1), P_{AE} for P_1, R_C for R_2, P_{CE} for f_1,

B for C(1 | 2), P_{AC} for P_2, R_E for R_1, and P_{CC} for f_2,

yielding:

$$ECM = A\,P_{AE} \int_{R_C} P_{CE}(x)\,dx + B\,P_{AC} \int_{R_E} P_{CC}(x)\,dx \tag{A5}$$

In our formulation these quantities represent the expected percentages of time that a record is classified as "erroneous" (R_E) or "correct" (R_C). Since the verification mechanism may be imperfect, records may be misclassified. That is, a correct record may be classified as erroneous, or vice versa. Noting that the first two terms of (A3) are identical to (A5), the stored MIS error rate may be defined as the expected percentage of time of misclassification plus a "remainder":

$$\text{Remainder} = P_{AE}\,E\left[\frac{\min(T_U + T_{C,E}, T)}{T}\right] + P_{AC}\,E\left[\frac{\min(T_U, T)}{T}\right] \tag{A6}$$

REFERENCES

Amer, T., A. D. Bailey, and P. De. 1987. A review of the computer information systems research related to accounting and auditing. *The Journal of Information Systems* 2 (No. 1): 3-28.

Ballou, D. P., and H. L. Pazer. 1985a. Modeling data and process quality in multi-input, multi-output information systems. *Management Science* 31 (No. 2): 150-162.

——— and ———. 1985b. Process improvement versus enhanced inspection in optimized systems. *International Journal of Production Research* 23 (No. 6): 1233-1245.

——— and ———. 1987a. Cost/quality tradeoffs for control procedures in information systems. *OMEGA: International Journal of Management Science* 15 (No. 6): 509-521.

——— and ———. 1987b. Designing inspection strategies for uncertain environments. *Decision Sciences* 18 (No. 2): 217-233.

———, ———, S. Belardo, and B. Klein. 1987. Implication of data quality for spreadsheet analysis. *Data Base* 18 (No. 3): 13-19.

Brodie, M. L. 1980. Data quality in information systems. *Information and Management* 3: 245-258.

Cooper, R. B. 1983. Decision production—a step toward a theory of managerial information requirements. *Proceedings of the Fourth International Conference on Information Systems.* Houston, TX. 215-268.

———. 1985. Identifying appropriate MIS/DSS support: A cost analysis approach. *Proceedings of the Sixth International Conference on Information Systems.* Indianapolis, IN. 89-104.

Cushing, B. E. 1974. A mathematical approach to the analysis and design of internal control systems. *The Accounting Review* 49 (No. 1): 24-41.

Emery, J. C. 1969. *Organizational Planning and Control Systems: Theory and Technology.* New York, NY: Macmillan.

Fellegi, I. P., and D. Holt. 1976. A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association* 71 (No. 353): 17-35.

Fields, K. T., H. Sami, and G. E. Sumners. 1986. Quantification of the auditor's evaluation of internal control in data base systems. *The Journal of Information Systems* 1 (No. 1): 24-77.

Gallegos, F., D. R. Richardson, and A. F. Borthick. 1987. *Audit and Controls of Information Systems.* Cincinnati, OH: South-Western Publishing Co.

Garfinkel, R. S., A. S. Kunnathur, and G. E. Liepens. 1986. Optimal imputation of erroneous data: Categorical data, general edits. *Operations Research* 34 (September-October): 744-751.

Groomer, S. M., and U. S. Murthy. 1989. Continuous auditing of database applications: An embedded audit module approach. *The Journal of Information Systems* 3 (No. 2): 53-69.

Hamlen, S. S. 1980. A chance-constrained mixed integer programming model for internal control design. *The Accounting Review* 55 (No. 4): 578-593.

Jaro, M. A. 1985. Current record linkage research. *Proceedings of the American Statistical Association*: 140-143.

Johnson, R. A., and D. W. Wichern. 1988. *Applied Multivariate Statistical Analysis*. 2d ed. Englewood Cliffs, CA: Prentice-Hall.

Johnson, J. R., R. A. Leitch, and J. Neter. 1981. Characteristics of errors in accounts receivable and inventory audits. *The Accounting Review* 56 (No. 2): 270-293.

Juran, J. M. and F. M. Gryna, Jr. 1980. *Quality Planning and Analysis*. New York, NY: McGraw-Hill.

Laudon, K. C. 1986. Data quality and due process in large interorganizational record systems. *Communications of the ACM* 29 (No. 1): 4-11.

Liepens, G. E., R. S. Garfinkel, and A. S. Kunnathur. 1982. Error localization for erroneous data: A survey. *TIMS/Studies in the Management Sciences* 19: 205-219.

Little, R. J. A., and P. J. Smith. 1987. Editing and imputation for quantitative survey data. *Journal of the American Statistical Association* 82 (No. 397): 58-68.

McKeown, P. G. 1984. Editing of continuous survey data. *SIAM Journal of Scientific and Statistical Computing*: 784-797.

Menkus, B. 1983. The problem with "errors." *Journal of Systems Management* 34 (No. 11): 11-13.

Morey, Richard C. 1982. Estimating and improving the quality of information in a MIS. *Communications of the ACM* 25 (No. 5): 337-342.

Nichols, D. R. 1987. A model of auditor's preliminary evaluations of internal control from audit data. *The Accounting Review* 62 (January): 183-190.

Soderlund, P. H. 1986. Controlling database costs through statistics. *Journal of Systems Management* 37 (No. 4): 26-31.

Stratton, W. O. 1981. Accounting systems: The reliability approach to internal control evaluation. *Decision Sciences* 12 (No. 1): 51-67.

Sumanth, D. J. 1984. *Productivity Engineering and Management*. New York, NY: McGraw-Hill.

Volonino, L., and P. Kirs. 1988. An investigation of performance, productivity, and rationality in multi-criteria decision making. *Proceedings of the Twenty-First Annual Hawaii International Conference on System Sciences* 3: 10-18.

Wand, Y., and R. Weber. 1989. A model of control and audit procedure change in evolving data processing systems. *The Accounting Review* 64 (January): 87-107.

Yu, S., and J. Neter. 1973. A stochastic model of the internal control system. *Journal of Accounting Research* 11 (No. 3): 273-295.