Journal of International Information Management

Volume 4 | Issue 1 Article 4

1995

Developing an optimization model for two-operator data entry

Roger G. Nibler
Lingnan College

Developing an optimization model for two-operator data entry

Roger G. Nibler
Lingnan College

ABSTRACT

Firms sometimes use two-operator data entry as a method to achieve or maintain database quality. When in-house staff are used, the firm typically selects data entry operators from a pool of junior staff and then assigns them into operator pairs, often on a random basis. Keying discrepancies between operator pairs are compared to determine incorrect entries, in the same row and column. Because the likelihood of making an error on a given key varies among operators, the objective of this study was to optimize database quality by systematically matching operators. The model was developed by having 32 operators key data into two databases and monitoring the location of each operator error. The database quality of all operator combinations were compared to determine optimal operator pairings. This resulted in 319% fewer errors in a second database over the expected number of errors which would have occurred from random operator pairings, and produced a database that was nearly 99.95% error-free. Regression analysis used operator error rates and total number of errors from each operator pair as independent variables. The dependent variable was the number of errors committed by each operator pair in the second database. The model explained 69% of the variability, and was used in a subsequent study using 28 different operators who entered a second database which was different from the first study. This resulted in a 0.85 correlation between the predicted versus actual in the second group.

INTRODUCTION

As firms become increasingly dependent on their Management Information Systems, their need for higher quality databases also increases. This necessity is a result of ongoing trends in the data processing field. One trend pertains to the fact that a given piece of information in a database is now more likely to be used by several subunits within the organization. For example, a Manufacturing Resources Planning II system coordinates sales, purchasing, manufacturing, finance, and engineering by adopting a central manufacturing plan and uses a single unified database to plan and update the activities in these systems (Adam & Ebert, 1992). Another trend pertains to liability issues. To illustrate, the Texas Court of Appeals recently held an oil pipeline company liable for negligently causing the delivery of 93,000 barrels of crude oil to the wrong consignee. The pipeline company was found liable for 50% of the damages for negligence, which arose from a data entry error (Westermeier, 1993). Because of these factors firms might be encouraged to develop a zero defects program for data entry.

During the past ten to fifteen years a number of changes have taken place that affect the data entry process of many companies. One change is distributed data processing, stemming from the development of PCs and Local Area Networks. This has been an important factor resulting in data being entered in a decentralized fashion as opposed to centralized systems using professional "Heads-down" data entry operators. Although this trend places data entry close to the source and therefore improves the opportunity for on-line validity checks, it can also result in the deterioration of database quality because data entry is often performed by a junior staff on a part-time basis in conjunction with other (and usually more interesting) duties. Moreover, in a number of cases, the data is entered by only one person, as opposed to the verifier method in which a second operator enters the same database and discrepancies between the two operators are compared and corrected.

Another change involves the data entry method. Recently there have been significant developments in source data automation such as Bar Coding, Image Recognition, Optical Character Recognition, Voice Recognition, and Electronic Pens. Many of these methods have improved database quality under certain conditions, but they continue to have their limitations. With automated entry, the input documents must meet higher clarity standards than for keyboard data entry (Gurney, 1992; Betts, 1991; Francis, 1991). In some situations the data to be entered must pass certain validity checks, which can only be done on a cost-effective basis by someone who is familiar with the data. In any event, the ultimate data source for most automated entry continues to be the human aided by mechanical or electronic devices. To this extent, automated data entry usually cannot be more accurate than the person who initially captured the data.

To raise or maintain the quality of critical databases, some companies should consider establishing minimum error standards for each database and then developing appropriate control measures. However, data entry is expensive and accounts for nearly 20% of the data processing costs in many firms (Bodeck, 1988; Rhodes, 1987). The challenge is to locate qualified staff to perform data entry, particularly in situations where data entry is spread throughout the functional areas of the firm and in dispersed geographical locations. A possible alternative is to use a professional data entry firm, especially one located offshore where labor is relatively inexpensive and savings of 25% - 50% can be obtained (O'Connor, 1992; Nurse, 1992; Anthes, 1991). Unfortunately, unless costly and sophisticated telecommunications equipment is utilized, long response times, loss of validity checks, and data confidentiality may negate cost savings. Since some firms wish to achieve and/or maintain database quality at professional data entry standards, which are often at the 99.95% level or higher, two-operator data entry using in-house staff may be the most cost effective alternative.

PURPOSE

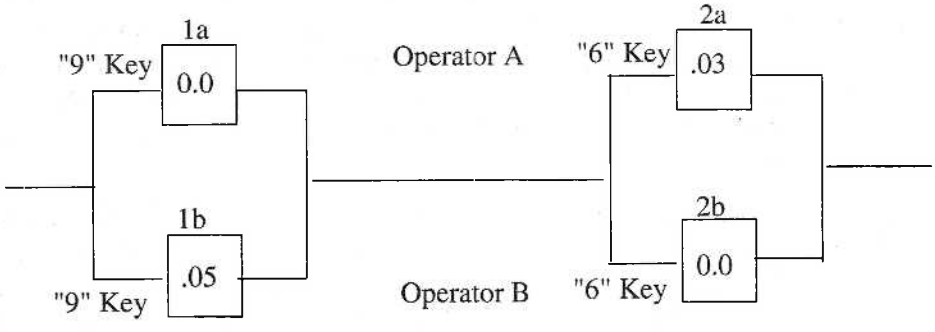
The objective of this study was to develop a model to optimize and predict database quality using two-operator data entry. One concept used in developing this model centers on a commonly held observation regarding keyboard data entry. In examining the nature and type of keyboard errors, the author and practitioners in the field have observed that operators

consistently tend to make the same types of errors even while their total error rate progressively decreases. For example, some operators tend to type a "9" when they mean to type a "6" because these two keys are close together and are struck by the weaker fingers when using a ten-key pad. Other operators may not tend to make this type of error, but may be more inclined to key an "a" when they mean to key an "s."

The second concept involves reliability theory of simple systems, whose outcomes are the joint probability of mutually exclusive events. To illustrate, if the probability of flipping a coin and having it come up heads is 0.5, then the probability of flipping a coin twice and having it come up heads each time is 0.25 [(0.5)(0.5) = 0.25]. In a similar manner, if system A has an error rate of 0.5 and system B operating in parallel with system A has an error rate of 0.3, then the probability that both systems would be in error at the same time is 0.15 [(0.5)(0.3) = 0.15]. The reliability/quality of the system would be 85% [1 - (0.5)(0.3) = 0.85] (Wadsworth, 1986).

The above reliability concept is an illustration which assumes that the subcomponents of the two systems have random error distributions. For a more complex example, assume that the above parallel systems each consisted of two subcomponents in series called 1a, 2a, for system A and 1b, 2b, for system B. If subcomponent 1a had an error rate of zero while subcomponent 2a had an error rate of 0.05, the overall error rate of system A would still be 0.05. Additionally, if subcomponent 1b had an error rate of 0.03 while subcomponent 2b had an error rate of zero, then the overall error rate of the total system would now be zero instead of 0.015 because any subsystem that could be in error is matched by a back-up system with zero probability of error. This is illustrated in Figure 1 below.

Figure 1. Parallel and Series Systems with Zero Defects Output



Reliability/Quality = [1 - (0.0) (.05)] [1 - (.03) (0.0)] = 100%
Error Rate = 1 - Reliability = 1 - 1.0 = 0.0%

A data entry operator can be viewed as a system consisting of approximately 60 different subsystems with each key representing a subsystem usually following a unique error distribution pattern. These subsystems can be considered linked in a series of various combinations to form a data entry session. Two-operator data entry would constitute a parallel system each composed of identical subsystems operating in parallel with each other, as illustrated in Figure 1.

If two operators enter the same data and discrepancies between them are examined to determine which operator is in error, then the simplified reliability model alone might be a highly accurate predictor of database quality if the operators had the same error patterns for each key during the data entry session. However, it is likely that operators will have different error patterns for a given key. Thus, some operator combinations are more likely to enter a database with fewer errors than other operator pairs, even in situations in which all the operators had the same error rate. In any event, an error always enters a database when the two operators make an identical mistake in the same row and column.

METHODOLOGY

The subjects used to develop the model were 32 entry level staff from a variety of medium sized companies in Hong Kong which tend to be data entry intensive, such as banks, insurance, and manufacturing. For these operators, approximately 15 to 30 percent of their work involved data entry. Thirty-two subjects were chosen because this is about the number of data entry operators used by these firms.

The operators keyed the input document against a known database, thus making it possible to record each operator error. The data to be input was that typically used by these companies and contained special words, abbreviations, symbols, and tables of numbers. The data to be keyed was displayed on the screen. This method of keyboard data entry is becoming increasingly popular as many companies are using scanners on input documents and the data to be keyed appears on the screen in image format.

When an operator made an error, the row, column, correct, and incorrect entries were recorded in an error database attached to the data entry software. The productivity, measured in keystrokes per hour, was also recorded for each operator. Validation software such as spell-checker was not used to control for the fact that data entry for some types of databases, especially those that are number-intensive and/or use special symbols, could not apply such software on a cost-effective basis.

The operators keyed the first database, which consisted of about 10,000 keystrokes. Then they returned in 6 hours to enter a second database, which was from a different source but similar in nature and length to the first. The second database provided a means to check operator data entry error consistency and served as a validation database to verify the optimization method for selecting operator pairs.

The operators were selected on a voluntary basis and were paid their approximate hourly wages to participate in the study. They were instructed to enter the data at a comfortable pace.

Other conditions such as the lighting and other environmental aspects were similar to their work environment.

Upon completion of this part of the study, the errors generated by each operator were assembled into an error database using software especially designed and programmed for the study by the author. The performance of each operator was compared against the remaining 31 operators such that all combinations of operators were examined to form a database containing 496 rows. This composite error database was then sorted into ascending order according to the number of errors committed by each operator pair in the first database. The first 15 rows of this database are shown in Table 1 below.

Table 1. Partial Error Database for All Operator Pairs

Row #	Errors (2nd) 1	Errors (1st) 2	Operator-1 Error Rate 3	Operator-2 Error Rate 4
*1	5.31	1.03	346	334
2	4.24	2.06	336	346
*3	3.18	2.06	389	282
4	3.18	2.06	345	520
*5	7.43	3.09	221	336
6	5.31	3.09	336	308
7	5.31	3.09	334	361
*8	4.24	3.09	226	381
9	1.06	3.09	381	282
10	2.12	4.13	226	221
11	6.37	4.13	221	389
12	6.37	4.13	346	389
*13	5.31	4.13	705	238
14	2.12	4.13	381	361
15	11.67	4.13	334	530

Errors (2nd): Errors/10,000 keystrokes in second database
Errors (1st): Errors/10,000 keystrokes in first database
* Unique Operator Pair

First Database: Average operator error rate: 466, Std. Dev.: 232
Second Database: Average operator error rate: 465, Std. Dev.: 300

After the model was developed, it was evaluated for its predictability by conducting a subsequent study using a different set of operators. These operators used the same first database as the previous group to establish their parameters for the model. Following that, they entered a different second database. A correlation analysis was then conducted to determine the degree of fit between actual errors in the second database and the number of errors predicted by the model.

RESULTS

Prior to developing the model, it was necessary to examine the consistency of the operators with respect to the errors they made in each of the two databases. To evaluate the error rate consistency for each operator between the two databases, the difference between matched pairs t-test resulted in a t value of 0.226 to yield p-value of 0.823. The mean and standard deviation of the operator errors for the first database was 466 and 232, while the second database was 465 and 300. Thus, the error rates averaged 4.66% with a standard deviation of 2.32% and a range of 2.21% - 15.24%. The high p-value indicates that the two databases were quite similar with respect to the number of errors committed by each operator. The productivity followed a similar pattern and averaged about 10,000 keystrokes per hour for each of the two databases with a standard deviation of about 3050 keystrokes per hour. An interesting finding was the inverse relationship ($r = -.61$) between productivity and the error rate. Generally, operators who had higher productivity also tended to have lower error rates.

The next step in developing the model was to examine the relationship between the theoretical and actual error rates for each of the two databases. The theoretical error rate is the product of the error rates of each operator pair to arrive at what the error rate for the database would have been according to the reliability concept.

Considering the previously mentioned likelihood of different error distributions for each key by a given operator, the overall actual error rate was tested to see if it was lower than the overall theoretical error rate for all possible pairwise operator combinations. For example, if operators A and B had error rates per 10,000 keystrokes of 320 and 250 respectively, their theoretical error rate would be 8 [(320)(250/10000 = 8]. However, if the actual error rate for these two operators was 2, then this would indicate that these two operators had significantly different error patterns on at least some of the keys. This was, indeed, the prevailing situation among all combinations of operators. The actual number of errors occurring in the database was less than that predicted by the theoretical error rate. These results are summarized in Table 2 on the next page.

Table 2. Relationship Between Theoretical and Actual Error Rates

	First Database Entered	Second Database Entered
	Actual Number of Errors	
Mean:	18.25	16.26
Std. Dev.:	10.89	19.73
	Theoretical Number of Errors	
Mean:	21.53	21.37
Std. Dev.:	14.91	19.73
	t-Test (one tail)	
D.F. = 495		
t Statistic:	-3.96	-4.74
p value	< 0.01	< 0.01

Although the actual errors were less than what the theoretical number of errors would suggest, this situation did not prevail among all operators. Out of 496 possible operator combinations, 52% had a lower number of actual errors as than theoretical errors for both databases, while 21% of the combinations contained operators whose actual errors were higher than their theoretical errors in both databases. The latter is indicative of operator pairs who tend to make the same type of errors, and for this reason should not be matched together for two-operator data entry. Approximately 27% of the operator combinations were not consistent in this regard between the two databases, with about an equal number of operators having more actual errors than theoretical errors in the first database than the second and vice versa.

The correlations between the actual and theoretical errors for each database were examined to determine whether regression analysis would be appropriate in developing a predictive model for database quality. These relationships are presented in Table 3.

Table 3. Correlations Between Actual and Theoretical Errors
N = 496

	1	2	3	4
1. Actual Errors (2nd Database)	*	.81	.79	.72
2. Actual Errors (1st Database)		*	.72	.71
3. Theoretical Errors (2nd Database)			*	.96
4. Theoretical Errors (1st Database)				*

These rather strong correlations can be attributed, in part, to the high degree of consistency not only in the error rates, but also the general stability of error distributions for a given key by a particular operator.

Based on these correlations, regression analysis was considered applicable, with the Actual Errors (2nd) being the dependent variable and combinations of the Theoretical (1st) and Actual Errors (1st) serving as the independent variables. The regression of the Actual Errors (1st) against the dependent variable was significant at the 0.0001 level (t = 30.24) and yielded an R-squared of 0.65. Regressing the Theoretical (1st) against the dependent variable was also significant at the 0.0001 level (t = 23.35) and yielded an R-squared of 0.52. Both independent variables together made the best contribution to the model. These results are presented in Table 4.

Table 4. Regression Analysis on Theoretical and Actual Errors

Dependent Variable: Actual Errors in Second Database

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	64441	32221	551	0.001
Error	493	28844	59		
C Total	495	93286			

R-Square: 0.69
Co-Variance: 47.04

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for HO: Param = 0	Prob > |T|
Intercept	1	-3.214	0.681	-4.723	0.0001
Actual Errors (1st)	1	0.746	0.046	16.279	0.0001
Theoretical (1st)	1	0.272	0.033	8.139	0.0001

Standard Error of Predicted Variable: 0.512

Because the correlation between these two independent variables was .72 there could potentially be a colinearity problem. However, the relatively lower covariance value of 47.04 compares favorably with the covariance values obtained when each of these independent variables were regressed separately, in which the Theoretical (1st) had a covariance value of 58.28 and the Actual Errors (1st) had a covariance value of 50.06. In addition, the R-Squared on the regression of these two independent variables was 0.69, which is not considered to be too high for multicolinearity consideration. Because of these two factors, multicolinearity does not appear to be a serious problem in this situation (Lardaro, 1993).

The regression explained 69% of the variability, and provided a strong indication that the model could have good predictability upon completion of the second step of the study.

A simplified model to optimize selection of operator pairs was developed which does not depend on parameters calculated from the second database, and thus does not encounter the problem of self-validation. This model was constructed by sorting the error database into ascending order according to Actual Errors (1st). In working down this sorted database, the first 5 unique pairs of operators resulting in the lowest total Actual Errors (1st) were selected. These operator pairs are indicated with an asterisk in Table 1. Five pairs were chosen because this number was considered likely to simulate a selection procedure for in-house data entry in which about one third of the eligible operators would be selected. In examining the performance of these particular operator combinations, it was noted that the Theoretical (1st) of all 5 operator pairs were greater than their Actual Errors (1st) by a multiple of 5.16. In observing the corresponding Actual Errors (2nd), their mean was 5.09 with a standard deviation of 1.41. This translates to a database quality of nearly 99.95% (1 - 5.09/10000), and contrasts rather favorably with the overall mean Actual Errors (2nd) of 16.26 with a standard deviation of 13.71. In this case, the mean and standard deviation of the Actual Errors (2nd) represents the expected value of what might occur if the selection of operator pairs was done on a random basis. Basically, this matching process resulted in a 319% decrease (16.26/5.09) in error rate over the expected value. As an additional step, these 10 operators were matched in various pair-wise combinations to determine their worst possible fit among each other. This would have resulted in an error rate of 20.66. Thus, it is important that the optimal matchings be maintained throughout the data entry process. Although the error rate of these operators was below average of all 32 operators, it was surprisingly close, with a mean and standard deviation of 348 and 133, respectively.

As a final step in validating the output of the simplified model, the error database was sorted into ascending order by Actual Errors (2nd) to determine the after-the-fact optimal arrangement of 5 operator pairs. Six of these operators were the same as those obtained by the selection model; however, none of the pairings were the same. The error rate from this group was 1.29, which would have been a 1260% improvement over the expected value. However, their worst possible fit would have resulted in an error rate of 16.31.

To test the predictability of the regression model, the SameMutual (1) errors and the Theoretical errors of the second set of 28 different operators were applied to the parameters of the regression model developed by the initial study using 32 operators. A correlation analysis between the predicted and actual error rate for all (378) operator pairs of this group was 0.85,

and thus explained 72% of the predicted versus actual variability. To this extent, it appears as though the parameters of this model may have good potential for application.

DISCUSSION

The results of this study indicate that selection techniques for determining optimum operator pairs may be enhanced by having operators enter a database and monitoring the errors made by each operator. Operators who had lower error rates also tended to make fewer errors, as intuition might suggest. However, the output of the simplified selection model as well as the much higher regression coefficient of 0.75 for the Actual Errors (1st) versus the weighting of 0.27 for the Theoretical (1st) suggests that operator matching can result in a further reduction in database errors. Moreover, the correlation of 0.85 on the actual versus predicted from the second study point to a potential application as an operator selection tool as well as a method to predict database quality.

Although the findings of this study pertain to data entry using human operators, potential application can be expanded to other parallel systems constructed from series component subsystems possessing varying but predictable error/reliability distributions. Such systems might include certain automated devices, manufacturing of critical components, and software development. Through optimal matching it may be possible to obtain better than expected overall system reliability, provided that an effective switching system(s) exists to conduct the transfer from defective to non-defective subsystems. In data entry, the switching system would be a human operator examining entry discrepancies against the input document, and therefore would be less than perfect. However, according to practitioners in the professional data entry field, this type of switching system typically has an error rate of less than .0001.

Unquestionably, further research is needed to expand and refine the results across a wider range of data entry conditions, especially with operators whose error rates fall outside the range used in this study. This study was conducted at the macro level in which overall error patterns among operators were used to develop the operator selection model. It is anticipated that a micro analysis which analyzes specific error patterns among the keys for each operator may yield insights which can refine the model.

ACKNOWLEDGMENTS

The author is most grateful to the information and support provided by Zhang Chi of KPT Software Ltd. and Kong Zhi Wei and Peter Wang of International Business Data Ltd. in Zhuhai, P. R. China. Also, a note of appreciation is due to the operators who devoted their time to enter data.

REFERENCES

Adam, E. E. & Ebert, R. J. (1992). *Production and operations management* (5th ed.). New Jersey: Prentice-Hall.

Anthes, G. H. (1991, August 26). U. S. firms go offshore for cheap DP. *Computerworld, 25*(34), 59-60.

Betts, M. (1991, September 8). IS must get a handle on bar coding. *Computerworld, 25*(36), 57-61.

Bodeck, N. (1988, June). DEMA ninth annual member statistical compensation survey. *DEMA: The Data Entry Management Association Newsletter.*

Francis, R. (1991, September 15). OCR comes down to the desktop. *Datamation, 37*(18), 44-46.

Gurney, B. & Discenza, R. (1992). Facilitating shop-floor bar-coding implementation: A do-it-yourself approach for small firms. *Production & Inventory Management Journal, 33*(4), 1-5.

Lardaro, L. (1993). *Applied econometrics*, 444-454. New York: Harper-Collins.

Nurse, L. (1992). Barbados: An action center for information services. *Telemarketing*, 78-80.

O'Connor, P. J. (1992, April). Outsourcing: It deserves a look. *The Office*, 14-16.

Rhodes, W. L. Jr. (1987, September). Data input enters a new era. *Infosystems*, 194-197.

Wadsworth, H.M., Stephens, K. S., & Godfrey, A. B. (1986). *Modern methods for quality control and improvement*, 620-654. New York: John Wiley & Sons.

Westermeier, J. T. (1993, Winter). Legal liability for insufficient error controls. *Information Strategy: The Executive's Journal, 9*(2), 54-55.