8-25-1995

Base Rates and Payoffs in the Detection of Errors in Data

Barbara D. Klein
University of Minnesota

Base Rates and Payoffs in the Detection of Errors in Data

Barbara D. Klein
Carlson School of Management
University of Minnesota
271 19th Avenue South
Minneapolis, MN 55455
(612) 624-8030

ABSTRACT. Information systems provide data to support business processes and decision making. There is strong evidence that data stored in organizational databases are not entirely accurate. If undetected, errors in data may significantly affect business outcomes. The goal of this research is to improve our understanding of the conditions under which individuals detect errors in data. The general motivating question is whether approaches to the management of data quality that depend upon users detecting errors are viable. Published research claims users are ineffective in detecting data errors. However, an exploratory study conducted as the first step in this research program suggests that users in one business profession (actuarial science) do detect errors in data. These conflicting findings motivate the specific research questions. This study initiates a stream of research on error detection by users of information systems through a laboratory experiment and a field study. The results are applicable to the design and management of information systems and provide the foundation for future research.

BACKGROUND. There is strong evidence (e.g., Laudon, 1986; Morey, 1982; Redman, 1992) that data stored in organizational databases are neither entirely accurate nor complete. Existing research suggests that humans are poor detectors of errors in these databases. For example, Laudon (1986) found that users do not verify the accuracy of data distributed by the FBI, Davis et al. (1967) found that half of the subjects in a field experiment failed to detect an error in a banking statement, and Ricketts (1990) found that ninety percent of the subjects in a laboratory experiment failed to detect an error in which expenses were inflated by at least a power of ten.

An exploratory study conducted as the initial phase of this research suggests that these earlier findings and generalizations drawn from them are at least partially incorrect. Interviews with eleven actuaries provide evidence that some users of data successfully detect data errors.

Two main approaches to the management of data quality are (1) validating data as they are input to or stored in databases (e.g., Morey, 1982; Ballou and Pazer, 1985; and Ballou et al., 1987) and (2) depending on users to detect and correct errors. Extensive validation is expensive, imperfect, and sometimes impractical. Still, if it is true that users are ineffective detectors of errors, investments in extensive data validation may be justified.

This research conceptualizes the detection of errors in data as an instance of fault diagnosis. A research model has been developed using signal detection theory (Green and Swets, 1966), Campbell's (1990) theory of individual task performance, and theories of adaptive decision making (Payne et al., 1982; Cryer et al., 1990). Signal detection theory is used to model performance. Signal detection theory suggests that performance measures should capture both the number of errors correctly detected and the number of accurate data values incorrectly labeled as errors. The findings of the study of actuaries, together with Campbell's theory of individual performance and theories of adaptive decision making, suggest that the base rate of errors, assessments of the payoffs of error detection, and users' goals will influence performance. These factors were used to identify variables to be manipulated in the laboratory experiment and to select three professional domains for the field study.

OBJECTIVES AND PROPOSITIONS. The goal of this research is to improve our understanding of the conditions under which individuals detect errors in data. An understanding of the conditions under which users do and do not detect data errors is important to the question of how information systems managers can best allocate resources to data quality efforts. This goal is met by two complementary studies: a laboratory experiment and a field study. Three propositions are tested in the study.

> Proposition 1. Expectations about the base rate of errors in data will affect performance in the detection of errors in data.
>
> Proposition 2. Payoffs will affect performance in the detection of errors in data.
>
> Proposition 3. Explicit error detection goals will affect performance in the detection of errors in data.

DESIGN AND METHODOLOGY.

Laboratory Experiment. The first study was conducted as a laboratory experiment examining the impact of expectations about the base rate of errors, incentive structures, and goals on error detection. The experimental task is from the domain of employee benefits and is based on incidents described by subjects in the study of the use of data by actuaries. The task requires subjects to use personnel data to calculate the accrued pension benefit for the employees of a set of six firms. Errors were embedded in the data based on an analysis of the task in consultation with a domain expert.

There are three independent variables: 1. expectations about the base rate of errors (high, low, or control), 2. incentive structure ($100 prize awarded based on overall task performance, $100 prize awarded based on the number of errors detected, $100 prize based on the number of errors detected with a penalty for incorrectly labeling accurate values as errors, or $100 prize awarded randomly), and 3. goal structure (explicit or implicit directions to review data for errors). Subjects were randomly assigned to levels of each factor. Theories of adaptive decision making and Campbell's theory of

performance predict that these factors will affect the number of errors detected and the number of accurate data values incorrectly identified as errors.

Dependent variables are the proportion of errors detected and the proportion of accurate values incorrectly identified as errors. Signal detection theory also suggests a measure of discriminability (the ability to distinguish accurate from inaccurate data) which will be compared across experimental conditions.

270 subjects recruited from upper-level undergraduate and graduate-level business courses participated in the experiment. Students were used because, compared to experienced professionals, they have weaker *a priori* expectations about the base rate of errors in data. The use of experienced professionals would have confounded the manipulation of the expectations about the base rate of errors factor because professionals may have strong *a priori* expectations developed through business experience.

Field Study. The second study is a replication of the field study examining the use of data by actuaries in three additional professional domains (consumer product management, municipal bond analysis, and inventory management). These domains were selected because they were expected to vary with respect to the base rate of errors and perceived payoffs of error detection. Theories of adaptive decision making and Campbell's theory of individual performance suggest that these domain differences will affect error detection. Structured interviews designed to elicit descriptions of incidents in which errors were detected were conducted with five professionals in each domain. The interview protocol is based on the research model of error detection. The interview transcripts were coded by two independent researchers using a list of codes which was developed from the constructs in the research model. Domain differences in the quantity and content of reported incidents will be identified. Differences will be used to analyze the extent to which the experimental findings are generalizable to organizations.

CONCLUSION. Errors in information processing have a significant impact on decision making in organizations. Because of the relative paucity of research in this area, questions which organizations face as they address data quality issues often cannot be informed by reference to research findings. The findings of this study have the potential for significantly improving the management of data quality in organizations by partially filling this gap.

An understanding of the conditions under which users do and do not detect errors in data has implications for the design and management of information systems. The finding that users detect errors under some conditions will suggest that approaches to the management of data quality that depend upon users to detect errors are viable. This finding is also an important prerequisite to the development of prescriptive interventions to improve error detection. If it is not possible to find instances of error detection in laboratory or field settings, we may conclude that the development of interventions to improve the ability or motivation of humans to detect errors in databases may be extremely difficult and that alternatives need to be researched.

REFERENCES.

Ballou, D. P., & Pazer, H. L. (1985). Modeling data and process quality in multi-input, multi-output information systems. **Management Science**, **31**, 150-162.

Ballou, D. P., Pazer, H. L., Belardo, S., & Klein, B. (1987). Implications of data quality for spreadsheet analysis. **Data Base**, **18**(3), 13-19.

Campbell, J.P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M.D. Dunnette and L.M. Hough (Eds.), **Handbook of Industrial and Organizational Psychology** (2nd ed., Vol. 1, pp. 687-732). Palo Alto, CA: Consulting Psychologists Press, Inc.

Cryer, E.H., Bettman, J.R., & Payne, J.W. (1990). The impact of accuracy and effort feedback and goals on adaptive decision behavior. **Journal of Behavioral Decision Making**, **3** , 1-16.

Davis, G. B., Neter, J., & Palmer, R. R. (1967). An experimental study of audit confirmation. **Journal of Accountancy**, **123** (6), 36-44.

Green, D.M., & Swets, J.A. (1966). **Signal detection theory and psychophysics**. New York: Wiley.

Laudon, K.C. (1986). Data quality and due process in large interorganizational record systems. **Communications of the ACM**, **29**, 4-11.

Morey, R. C. (1982). Estimating and improving the quality of information in a MIS. **Communications of the ACM**, **25** , 337-342.

Payne, J.W. (1982). Contingent decision behavior. **Psychological Bulletin**, 92, 382-402.

Redman, T. C. (1992). **Data quality: Management and technology** . New York, Bantam Books.

Ricketts, J. A. (1990). Powers-of-ten information biases. **MIS Quarterly**, **14**, 62-77.