December 1998

User Detection of Errors in Data: Learning through Direct and Indirect Experiences

Barbara Klein
University of Michigan-Dearborn

User Detection of Errors in Data: Learning through Direct and Indirect Experience

Barbara D. Klein
University of Michigan-Dearborn

Abstract

Previous research shows that organizational databases contain data errors. If undetected, these data errors may significantly affect business outcomes. The study described in this paper examines two ways that users may learn to detect these data errors: direct experience and indirect experience. The results will have implications for the design of information system training programs.

Background

There is strong evidence (e.g., Laudon, 1986; Morey, 1982) that data stored in organizational databases are neither entirely accurate nor complete. If undetected, errors in data may significantly affect business outcomes. Two main approaches to this problem are (1) validating data as they are input to or stored in databases (e.g., Morey, 1982) and (2) depending on users to detect and correct errors.

A research program examining the efficacy of the second approach is underway. To date, a field study and a laboratory experiment have demonstrated that direct experience finding errors in data improves error detection performance. Field interviews were conducted with 25 professionals in four domains (actuarial science, consumer product management, inventory management, and municipal bond analysis) (Klein et al., 1996). The objective of the field study was to examine the relationship between expectations about the base rate of errors and error detection. Base rate expectations of the subjects interviewed in this study had been developed through several years of direct experience working with data. A relationship between expectations about the base rate of errors and self-reports of error detection was found ($p<.05$) (Klein et al., 1996).

An experiment was conducted to bolster our confidence about this relationship by studying the relationship between experience-based expectations about the base rate of errors and objective measures of error detection performance in a controlled laboratory setting. In this experiment, subjects learned about the error rate through direct experience. A relationship between the base rate of errors and error detection performance was found (Klein, 1996).

The laboratory experiment described in this paper builds on these prior studies and is designed to improve our understanding how users of information systems learn about data errors. The study is a first step toward understanding how training programs might improve user detection of errors. The study examines two ways that users might learn to detect data errors: direct experience and indirect experience.

Direct experience is defined as a past episode in which a user of data found a data error. Prior studies suggest that direct experience with data errors improves future error detection performance (Klein, 1996; Klein et al., 1996). However, learning about data errors through direct experience may be time consuming, expensive, and risky in some organizational settings. Also, in organizational settings with an extremely low error rate, opportunities for learning about data errors through direct experience may be very limited.

Learning about data errors through indirect experience may provide an alternative approach in organizational settings in which learning about data errors through direct experience is not appropriate. Indirect experience is defined as reading or hearing about a past episode in which another user of data found a data error. For example, a new employee might attend a training session or read a users' guide describing a number of data errors found in the past by employees doing similar work in the organization.

A research model guides the work in this research stream. The model uses signal detection theory (Green and Swets, 1966), Campbell's (1990) theory of individual task performance, theories of adaptive decision making (Payne, 1982), and theories of expertise (Ericsson and Chase, 1982).

Objectives and Propositions

The goal of this research is to improve our understanding of the conditions under which individuals detect errors in data. Five propositions are tested in the study.

Proposition 1: Direct experience with data errors influences performance in error detection. This proposition is based on Campbell's (1990) theory of individual performance and theories of expertise (Ericsson and Chase, 1982).

Proposition 2: Indirect experience with data errors influences performance in error detection. This proposition is based on Campbell's (1990) theory of individual performance.

Proposition 3: Indirect experience with data errors and direct experience with data errors have the same effect on performance in error detection.

Proposition 4: There is an interaction between direct experience with data errors and incentive structures. This proposition is based on signal detection theory (Green and Swets, 1966), Campbell's (1990) theory of

individual performance, and theories of adaptive decision making (Payne, 1982). The specific prediction is that incentives will be associated with greater differences in error detection performance between users with direct experience and with no experience.

Proposition 5: There is an interaction between indirect experience with data errors and incentive structures. This proposition is based on signal detection theory (Green and Swets, 1966), Campbell's (1990) theory of individual performance, and theories of adaptive decision making (Payne, 1982). The specific prediction is that incentives will be associated with greater differences in error detection performance between users with indirect experience and users with no experience.

Design and Methodology

The study is being conducted as a laboratory experiment. The experimental task is from the domain of employee benefits. The task and the experimental materials have been successfully used in prior studies within this research stream (Klein, 1996; Klein et al., 1997). Subjects use personnel data to calculate the pension benefit for the employees of two firms. Errors were embedded in the data based on an analysis of the task in consultation with a domain expert. Subjects have ninety minutes to perform the experimental task.

There are two independent variables: 1. experience (direct, indirect, or none) and 2. incentive structure ($100 prize awarded based on the number of errors detected or no prize). Subjects in the direct experience condition will perform the pension calculations using a dataset containing twenty errors before performing the pension calculations for a second target company. Only subjects finding five or more errors will be retained in the experimental analysis. Subjects finding fewer than five data errors will be omitted due to a lack of direct experience detecting data errors. Subjects in the indirect experience condition will read a description of five data errors before performing the pension calculations for the target company. The five errors will be representative of those in the first dataset used by the subjects in the direct experience condition. Subjects in the no experience condition will only perform the task for the target company. The dataset used for the target company will be identical for subjects assigned to all three levels of the experience condition. The levels of the incentive structure used in this study have been found to affect the detection of data errors in prior research (Klein, 1996; Klein et al., 1997).

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. Only performance on the target company will be measured.

160 subjects recruited from upper-level undergraduate and graduate-level business courses will participate in the experiment. This number of subjects allows as many as forty subjects in the direct experience condition to be omitted from the analysis if they do not detect at least five data errors in the first dataset.

A questionnaire will be used to check whether subjects' perceptions are consistent with the levels of the treatments to which they were exposed.

The effects of the experience factor and the incentive structure factor will be investigated using standard ANOVA models. If interaction or main effects are found, the Tukey method of multiple comparisons will be used to test for differences in performance between specific treatment groups.

Conclusion

The results of this investigation will improve our understanding of how user learning about data errors affects error detection performance. An understanding of the efficacy of learning about data errors through indirect experience has implications for the design of training programs for information systems. For example, classes and user documentation might be designed to include information about data errors found in the past. Users might also be encouraged to share information about data errors they find, perhaps by posting such information on a corporate Intranet.

References

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.

Ericsson, K. A., & Chase, W. G. (1982). Exceptional memory. American **Scientist**, **70**, 607-614.

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

Klein, B. D. (1996). Base rate expectations and the detection of errors in data: Direct experience versus information. **Proceedings of the Second America Conference on Information Systems**, pp. 197-199.

Klein, B. D., Goodhue, D. L., & Davis, G. B. (1996). Conditions for the detection of data errors in organizational settings: Preliminary results from a field study. Proceedings **of the 1996 Conference on Information Quality**, pp. 24-52.

Klein, B. D., Goodhue, D. L., & Davis, G. B. (1997). Can humans detect errors in data?: Impact of base rates, incentives, and goals. **MIS Quarterly**, **21**(2), pp. 169-194.

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.