JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods in Enterprises
Behavior Analysis under Risk an Uncertainty

DEVELOPMENTS IN ANALYSIS OF MULTIPLE RESPONSE SURVEY DATA IN CATEGORICAL DATA ANALYSIS: THE CASE OF ENTERPRISE SYSTEM IMPLEMENTATION IN LARGE NORTH AMERICAN FIRMS[1]

Kayvan Miri LAVASSANI[2]
PhD Candidate, Research Associate, Sprott School of Business,
Carleton University, Ottawa, Canada

E-mail: Kayvan@Lavassani.ca

Bahar MOVAHEDI[3]
PhD Candidate, Research Associate, Sprott School of Business,
Carleton University, Ottawa, Canada

E-mail: Bahar_Movahedi@Carleton.ca

Vinod KUMAR[4]
PhD, Professor, Technology & Operations Management, Sprott School of Business,
Carleton University, Ottawa, Canada

E-mail: Vinod_Kumar@Carleton.ca

Abstract: *This paper explores the analysis of survey data with multiple response variables. After describing the problem with analysis of multiple response variables, the historical developments in identifying and analyzing multiple response variables, based on an extensive literature review, are discussed. After we explored the developments in this area from 1968 to 2008, we employed the first Order Rao-Scott Corrected Chi-Square to analyze a recently collected set of data on the practice of Enterprise System (ES) implementation among North American large corporations. The data analyzes the success of ES implementation, challenges of ES implementation, and the success of utilization of ES across two categories of firms: process-oriented and not process-oriented. The first Order Rao-Scott Corrected Chi-Square confirms that process-oriented firms in our sample are more successful in implementing the ES, face fewer challenges in implementing the ES, and are more successful in utilizing the ES.*

Key words: *Complex Survey Data; Enterprise System Implementation; Empirical; Multiple Response Variable; Categorical Data*

1. Introduction

Analyzing complex data collected from the surveys is one of the challenges facing the researchers. The complexity of the data is a multifaceted issue and has different implications. One of these challenges (facets) comes when researchers working with categorical data are working with multiple response variables. This problem arises when, for a single observation, a variable or some variables may be classified into more than one category. We should note that the cause of this type of complexity is "the multiple-response nature of the data, not from the sampling mechanism" or the design of the questionnaire (Thomas and Decady, 2004). When more than one answer may be selected by the respondents, the response for a single observation can be classified into more than one category. The problem of multiple response variables can be observed and studied in n-way contingency tables. The focus of this study is on the problem of multiple response variables in two-way contingency tables, while the situation of Enterprise System (ES) implementation presents a case of single-by-multiple marginal independence. The research problem is explained in section two by presenting a generic example. The next section explores the historical developments in identifying and understanding the multiple response variables in categorical data analysis. Section four presents the application of new statistical tools in analyzing data recently collected from a sample of large North American firms; the data is examined to determine the success in implementing ES, the challenges of implementing ES, and the success of utilizing ES. Finally, section five presents the conclusion and gives suggestions for future studies.

2. The Problem with Multiple Response Variables in Categorical Data Analysis

The issue of multiple response variables is becoming more and more visible and, therefore, has attracted the attention of researchers and practitioners, specifically in the past decade. For example, in a recent guideline prepared by the Australian Institute of Health and Welfare for those involved in collecting and presenting the data regarding alcohol and other drug treatment, the issue of multiple response variables as an "indigenous status question" has been identified (Australian Institute of Health and Welfare working paper, 2008).

Although the existence of multiple response variables may be easily identified, the implication of analyzing multiple response variables has received less attention. There are numerous studies dealing with multiple response variables. However, in some cases, the researchers simply ignored the fact that when they are dealing with multiple response variables. Specifically the chi-square test is not a reliable test when multiple response variables are being analyzed. One example is the Stallings and Ferris (1988) study on public administration research where, despite the recognition of multiple response variables, the researchers have used the simple chi-square test to identify the difference between different categories of data. Decady and Thomas (2000) explicitly described two main reasons that the Pearson chi-square test is not appropriate in dealing with multiple response variables. Here we will describe the problem with multiple response variables using a generic example. Consider the 2x2 contingency table (Table 1). First, we assume that there are no multiple response variables.

Table 1. A 2-by-2 table of observations with no multiple response variables

	Y1	Y2	
X1	a_{11}	a_{12}	$a_{11}+a_{12}=N_{1+}$
X2	a_{21}	a_{22}	$a_{21}+a_{22}=N_{2+}$
	$a_{11}+a_{21}=N_{+1}$	$a_{12}+a_{22}=N_{+2}$	$=N=a_{++}$

In this table, the observed counts are presented in four cells. X is the independent variable and Y presents the response variable. The marginal values are presented by N_{+1}, N_{+2}, N_{1+}, and N_{2+}. In each row and column the marginal values present the summation of that row or column. The Pearson chi-square test is calculated by the following formula:

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \qquad (1)$$

We have the observed variables in Table 1. We also need the expected value of each observation, based on the marginal totals, for the ability to calculate the Pearson chi-square. Table 2 presents the way the expected values are calculated.

Table 2. A 2-by-2 table of expected values with no multiple response variables

	y1	y2	
x1	$(N_{1+} * N_{+1})/N$	$(N_{1+} * N_{+2})/N$	N_{1+}
x2	$(N_{2+} * N_{+1})/N$	$(N_{2+} * N_{+2})/N$	N_{2+}
	N_{+1}	N_{+2}	N

Here the two components of the Pearson chi-square are displayed: observed (Table 1) and expected (Table 2) values. However, this presentation is based on the assumption that none of the independent (e.g., rows) and response (e.g., columns) variables have multiple response variables.

Now if we assume that some variables can receive multiple responses for any row and/or column, then the marginal values of that row or column (there may be more than one of either) would be greater than the total observations of the variables. In this situation, the calculation of expected values using the model proposed above would be problematic. This is the first reason that Decady and Thomas (2000) gave when they stated that the traditional chi-square test is not appropriate for these circumstances. The second reason is that since one observation in this circumstance may yield multiple responses, the "standard assumption" of independence of rows and columns in he table is violated (Decady and Thomas, 2000). Further to these theoretical explanations, Rao and Scott (1981, 1984, and 1992) empirically showed that "classical chi-squared tests are invalid when applied to data from complex sample survey because the complexities of the survey design violate[s] the assumptions on which these tests are based" (Decady and Thomas, 2000).

3. Historical Developments in Analyzing Multiple Response Variables

Previously, two main reasons were given to explain why the classical Pearson chi-square is not appropriate for analyzing the complex survey data with multiple response variables. In this section of the paper, we explore how the researchers in academia deal with the analysis of the multiple response variables. To explore the evolution of studies in this

area, we conducted an extensive literature search. We used a number of academic databases to identify the data on the evolution of studies in this area. The following presents the result of our literature analysis.

The analysis of complex survey data has been of interest to researchers outside the field of mathematics and statistics since the 1970s. For example, Irving Roshwalb (1973) mentioned the "need [to] improvement" of analytical techniques for handling the complex survey data. In the 1980s, advancements were made by statisticians to provide more sophisticated analytical tools. For example, Fellegi (1980) focused on the tests of independence in complex samples. As mentioned previously, the complexity of sample data has different dimensions and the focus of this study is on the "multiple response variables," which is only one facet of complex survey data. It is not clear when exactly the problem of multiple response variables as a research topic and statistical problem was introduced. Our review of the literature in different domains showed that an early recognition of the attention to the multiple response variables came in 1968 in the work of Murphy and Tanenhaus (1968) in the U.S. Survey Research Center. In another study, Schriesheim et al. (1974) explored the development of response categories in the validity of multiple response alternative questionnaires. However, in these works, Murphy and Tanenhaus (1968) and Schriesheim et al. (1974) have provided no discussion regarding the data analysis; they gave basically a mention of the existence of the multiple responses due to the nature of the data. Not until the early 1980s did some statisticians publish papers specifically addressing this topic as a research issue.

The review of the studies in this area showed that some of the studies have simply ignored the problems with multiple response variables in analyzing categorical data. An example is the study of Stallings and Ferris (1988) on the two categories of policy and management topics in the journal of public administration review, which was mentioned previously. In this study, while Stallings and Ferris (1988) recognized the existence of multiple response variables, they used the classical (Pearson) chi-square in their analysis, which is not an appropriate tool (as explained previously) for analyzing such complex data. In some other studies where the collected data could lead to the issue of analyzing multiple response variables in some cases, the researchers preferred to change the method of collecting or analyzing the data in order to avoid dealing with multiple response variables in contingency tables. While this approach is effective in avoiding multiple response variables, in some cases it may lead to partial collection of data.

One of the early approaches in providing a tool for dealing with this problem was done by Umesh (1995). In his study, Umesh recommended the use of a modified pseudo-chi-squared test instead of the classical Pearson chi-square test. Umesh's recommendation was tested by Loughin and Scherer (1998) and the evidence showed that, under some conditions, this method fails to provide a strong control of test levels. In the late 1990s, Agresti and Liu (1998, 1999) advanced the understanding of the multiple response categorical variables. Furthermore, Loughin and Scherer (1998) proposed the use of the bootstrapping technique for estimating the p-value of their proposed statistic. This method attracted the attention of academia, where it was recommended that the Imhof (1961) methods of evaluating the probability density function (pdf) could also be used to estimate the p-value (Decady and Thomas, 2000). Further, scholars proposed solutions to continue exploring the application of bootstrapping in analyzing contingency tables with multiple response variables. For example, Bali et al. (2006) proposed a bootstrapping technique

considering the residuals of cells. While bootstrapping showed good control variables, Decady and Thomas (2000) tried to provide a simpler method that not only required less computation but also was more familiar to the practitioners. For achieving this goal, Decady and Thomas (2000) "cleverly draw the connection between the MMI (multiple marginal independence) testing problem and the Rao and Scott (1981) analyses of complex survey data" (Bilder and Loughin, 2001). They "note[d] the parallel between an application of an adjusted Pearson statistic to multiple-response categorical variables and the use of the Pearson statistic in non-multinomial sampling structures as studied by Rao and Scott (1981)" (Bilder and Loughin, 2007).

Although Bilder and Loughin (2001) recognized the contribution of the modified chi-square proposed by Decady and Thomas (2000), they questioned the control of the first order modified Decady-Thomas chi-square. In 2004, Thomas and Decady presented the extension of Rao and Scott modified chi-square, which was based on the second order Rao and Scott test. This recent procedure showed a good control of the test levels (Type I errors). More recently, Bilder and Loughin (2003, 2007) helped to further advance this area by exploring the extension to multiple-response categorical variables, which was originally proposed (but not conducted) by Agresti and Liu (1999, 2001).

4. The Case of ES Implementation: First Order Rao-Scott Corrected Chi-Square

In this paper, the first order Rao-Scott modified chi-square has been employed in a case of multiple response data recently collected from the survey of large North American corporations (V. Kumar et al., 2008; U. Kumar et al., 2008). In this empirical study, the authors measured the following four constructs of implementing ES:

- Process orientation
- Success of ES implementation
- Challenges during implementation of ES
- Successful utilization of ES

Each of these constructs is assessed by several measured constructs that are explicitly explained by the authors. Here is a brief description of the measured constructs.

ES in this survey is defined by the authors as an integrated, customized and packaged software based system that handles the majority of systems requirements in all or any of the functional areas of a firm, such as marketing, finance, human resources, and manufacturing. Almost every medium and large organization has at least a number of Enterprise Systems (ES) modules, such as Company-wide Accounting Software Package, Marketing Software Package or Manufacturing Software Package.

Furthermore, the concept of Process Orientation is described as "the activity of transforming an organization's structure from one based on a functional paradigm to one based on a process paradigm. Business process orientation implies that the procedure of doing tasks in firms should be more cooperative and integrated towards satisfying the customers' needs. This view is in contrast with the mechanistic functional view of the firm, which emphasizes the division and isolation of functions from each other and from the customers. While the challenges of ES include different dimensions of ES implementation, the concept of success is explored in two contexts: ES implementation and ES utilization. The questionnaire was sent to approximately 3,000 large North American firms. The survey

yielded a response rate of approximately 10 percent; 195 of the surveys were found to be complete enough to be used in a contingency table for the purpose of this study.

For analyzing these data, a 2x3 way contingency table was constructed (Table 3). For the construct of process orientation, each observation can only have a single response (whether process-oriented or not-process-oriented); for the other three constructs each observation can be multiple responses. In other words, in each observation the firm, whether process oriented or not, is actually process oriented. However, irrespective of its process orientation, a particular firm that was observed may:

- Be successful or unsuccessful in ES implementation,
- Face or not face significant challenges during ES implementation, and
- Be successful or unsuccessful in utilizing ES.

Table 3. Contingency Table of Constructs of ES Implementation

	Success in Implementation	Faced No Significant Challenge	Success in Utilization	Total Responses	Total Subjects
PO	88	100	101	**289**	**101**
Not-PO	71	36	77	**184**	**94**
	159	**136**	**178**	**473**	**195**

Marginal values in this contingency table (Table 3) clearly show the existence of multiple response variables in the data. In this case we are facing a single-by-multiple marginal independence.

4.1. First Order Rao-Scott Corrected Chi-Square

As described earlier, the use of traditional chi-square is not appropriate when dealing with multiple response data. Following Decady and Thomas (2000) in this study, a corrected Rao-Scott chi-square test will be applied. The corrected Rao-Scott chi-square test is presented as Equation 2:

$$X_C^2 = X^2/\tilde{\delta}, \tag{2}$$

Where:

χ^2_c Presents the Corrected Rao-Scott Chi-Square

χ^2 Presents the Traditional (Pearson) Chi-Square

$\tilde{\delta}$ Presents the Correction Factor

The correction factor ($\tilde{\delta}$) was calculated using Equation 3:

$$\tilde{\delta} = 1 - \frac{m_{++}}{n_+ C} \tag{3}$$

Where:

m_{++} Presents the total count of multiple responses, which here is equal to 473
n_+ Presents the total number of subjects, which here is equal to 195
C Presents the number of multiple response variables, which here is 3 (columns)

⇨ $\tilde{\delta}$ = 1-(159+136+178)/(195x3)= 0.1915

Additionally, the degree of freedom here is calculated as follow:
$(R-1)C_{d.f.}$

Where:

R Presents the number of rows related to the single response variable, which here is equal to 2

⇨d.f.= (2-1)*3=3

Now having $\tilde{\delta}$, d.f, and the (Pearson) chi-square, we can calculate the corrected Rao-Scott chi-square as follows:

$X^2 = 12.4774^5$ ⇨ $X^2_c = 12.4774/0.1915 = 65$ ⇨ *p*-value=0.000

Based on the corrected chi-square test, we have concluded that the process-oriented firms, in comparison to the not-process oriented firms:

- Are more successful in implementing ES
- Face fewer challenges in implementing ES, and
- Are more successful in utilizing ES.

It is important to note that the traditional chi-square test also showed almost similar results in the p-value (see footnote 1). This was due to the fact that differences between the two categories of process oriented and not-process oriented firms were significantly wide. However, it by no means justifies the use of traditional chi-square in this circumstance, as was described earlier.

5. Conclusion and Future Studies

In this study, one dimension of complex survey data – multiple response variables – was explicitly explored. The analysis of multiple response variables in contingency tables is a relatively (as compared to some other statistical research topics) new research problem. This study presented the historical developments of the studies in this area. In reviewing the historical developments of the complex research data and, specifically, the multiple response variables, several academic databases were employed.

The first order Rao-Scott chi-square was employed to analyze our data. The findings confirm that process-oriented firms in our sample – in comparison to the not-process oriented firms – were more successful in implementing ES, faced fewer challenges in implementing ES, and were more successful in utilizing ES. Furthermore, the first order Rao Scott corrected chi-square was employed to assess the results of the current survey data.

Table 4. Comparison of First and Second order Rao-Scott chi-square

	First order Rao-Scott chi-square	Second order Rao-Scott chi-square
General Formula (Simple-by-Multiple)	$X_C^2 = X^2/\tilde{\delta},$	$X^2_{SM}(AL)/(1+\hat{a}^2)$
Correction Factor	$\tilde{\delta} = 1 - \frac{m_{++}}{n_+ C}$	$\hat{a}_*$
Degree of Freedom	*df: (r – 1)c*	*df: (r – 1)c/(1 + â ^ 2)*

$*\hat{a}$ **:** is an estimate of the variability, among the weights that takes the form of a coefficient of variation

The comparison of first and second order Rao-Scott chi-square is displayed in Table 4. In future studies, the second order corrected Rao-Scott chi-square test could be employed.

References

1. Agresti, A. and Liu, I.M. **Strategies for modeling a categorical variable allowing multiple category choices**, Sociological Methods and Research, 29, 2001, pp. 403-434
2. Agresti, A. and Liu, I.M. **Modeling responses to a categorical variable allowing arbitrarily many category choices**, Technical Report 575, University of Florida, Department of Statistics, Gainesville, 1998
3. Agresti, A. and Liu, I.M. **Modeling a categorical variable allowing arbitrarily many category choices**, Biometrics, 55, 1999, pp. 936-943
4. Bali, G. Ch., Czerski, D., Klopotek, M. A. and Matuszewski, A. **Residuals for Two-Way Contingency Tables, Especially Those Computed for Multiresponces**, Advances in Soft Computing, 5, 2006, pp. 201–210
5. Bilder, C. R. and Loughin, T. M. **On the first-order Rao-Scott correction of the Umesh-Loughin-Scherer statistic**, Biometrics 57, 2001, pp. 1253-1255
6. Bilder, C. R. and Loughin, T. M. **Strategies for modeling two categorical variables with multiple category choices**, American Statistical Association Proceedings of the Section on Survey Research Methods, 2003, pp. 560-567
7. Bilder, C. R. and Loughin, T. M. **Modeling association between two or more categorical variables that allow for multiple category choices**, Communications in Statistics: Theory and Methods, 36, 2007, pp. 433-451
8. Bilder, C. R. and Loughin, T. M. **Estimation and Testing for Association with Multiple-Response Categorical Variables from Complex Surveys**, Proceedings of the Survey Research Methods Section, ASA, 2007
9. Decady, Y. J. and Thomas, D. R. **A simple test of association for contingency tables with multiple column responses**, Biometrics, 56, 2000, pp. 893-896
10. Fellegi, I. **Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples**, Journal of the American Statistical Association, Vol. 75, No. 370, 1980, pp. 261-268
11. Imhof, J. P. **Computing the distribution of quadratic forms in normal variables**, Biometrika, 48, 1961, pp. 419-426
12. Kumar, V., Lavassani, K. M., Kumar, U., and Movahedi, B. **An Exploration on the Role of Process Orientation in Enterprise Systems Implementation,** Proceedings of Production Operation Management Society, 2008, California, USA
13. Kumar, U., Movahedi, B., Kumar,V. and Lavassani, K. M. **Measurement of Business Process Orientation in Transitional Organizations: An empirical study**, in Abramowicz W. and Fensel D. (Eds.) "Business Information Systems", Berlin, Germany, Springer, 2008, pp. 357-368
14. Loughin, T. M. and Scherer, P. N. **Testing for association in contingency tables with multiple column responses**, Biometrics, 54, 1998, pp. 630-637

15. Murphy, W. F. and Tanenhaus, J. **Inter-university Consortium for Political and Social Research US Survey Research Center 1966 American National Election Study: Post-election Interview**, Nov. 9, 1966
16. Rao, J.N.K. and Scott, A.J. **On Chi- Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data**, Annals of Statistics, 12, 1984, pp. 46-60
17. Rao, J. N. K. and Scott, A. J. **A simple method for the analysis of clustered data**, Biometrics, 48, 1992, pp. 557-585
18. Rao, J. N. K. and Scott, A. J. **The Analysis of Categorical Data From Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables**, Journal of the American Statistical Association, Vol. 76, No. 374, Jun. 1981, pp. 221-230
19. Roshwalb, I. **Review: A Timely Festschrift Probing the State of the Arts in Behavioral Economics**, Journal of Marketing Research, Vol. 10, No. 3, 1973, pp. 354-356
20. Schriesheim, C. and Schriesheim, J. **Development and Empirical Verification of New Response Categories to Increase the Validity of Multiple Response Alternative Questionnaires**, Educational And Psychological Measurement, vol. 34, no. 4, December 1974, pp. 877-884
21. Stallings, R.A. and Ferris, J.M. **Public Administration Research: Work in PAR', 1940-1984**, Public Administration Review, Vol. 48, No. 1, Jan. - Feb., 1988, pp. 580-587
22. Thomas, D. R. and Decady, Y. J. **Testing for Association Using Multiple Response Survey Data: Approximate Procedures Based on the Rao-Scott Approach**, International Journal of Testing, Vol. 4, Issue 1, 2004, pp. 43 – 59
23. Umesh, U. N. **Predicting nominal variable relationships with multiple responses**, Journal of Forecasting, 14, 1995, pp. 585-596
24. * * * Australian Institute of Health and Welfare, **Alcohol and other drug treatment services NMDS specifications 2008–09**, Data dictionary, collection guidelines and validation processes. Working paper, 2008, URL: http://152.91.62.50/publications/hwp/hwp06/hwp06.pdf [Visited on June 1st, 2008]

---

[1] **Acknowledgement**
We extend our sincere appreciations to Professor Roland Thomas for his guidance with categorical data analysis.

[2] Kayvan Miri Lavassani is a Research Associate at the Carleton University. He has worked as a manager and an entrepreneur with high-tech, manufacturing and consulting firms in Canada and internationally. In the past few years Kayvan has published over 35 papers in refereed journals, book chapters, and conference proceedings. Kayvan has won several internal and external awards for his academic achievements. He is currently pursuing his Ph.D. program at the Sprott school of Business at Carleton University, Ottawa, Canada.

[3] Bahar Movahedi is a Research Associate at Carleton University, Ottawa, Canada. Movahedi has a MBA from Sprott School of Business at Carleton University, and is currently pursuing her Ph.D. program. Her main research areas are Technology Transfer and e-Commerce. In the past few years she has published over 35 papers in refereed journals, book chapters, and conference proceedings, and has won several internal and external awards for his academic achievements.

[4] Vinod Kumar a graduate of the University of California, Berkeley, is a Professor of Technology and Operations Management at the Sprott School of Business (Director of School, 1995-2005), Carleton University. He has published over 150 papers in refereed journals and proceedings. He has won several Best Paper Awards in prestigious conferences, Scholarly Achievement Award of Carleton University for the academic years 1985-1986 and 1987-1988, and Research Achievement Award for the year 1993, 2001 and 2007. He is on the editorial board of two international journals. He has also served for several years on the Board of Governors and the Senate for Carleton University.

[5] p-value $_{\chi 2}$=0.000