Decision Complacency, Consensus and Consistency in the Presence of Data Quality Information

InduShobha N. Chengalur-Smith
Harold L. Pazer
School of Business
SUNY-Albany
Albany, NY 12222

Abstract

The currently reported research is the third and final stage of a pilot study exploring the impact of data quality tagging on decision making. Our goal is to assist those who design databases to support decision processes by providing insight into what type of data quality information would be most effective. This study investigates the effect of individual differences as measured by differences in learning styles on three key measures: decision complacency, decision consensus and decision consistency. Learning styles were used as a surrogate for the mindsets of different professionals in an organization. The results show that there are no significant differences that are attributable to learning style. This makes for simpler implementation since this implies that the data quality information for a database does not have to be tailored to individual departments.

Introduction

It is a well-known fact that the past decade has seen an explosion in the number of databases designed primarily to support managerial decision processes. The quality of the data in such systems varies from intensely audited financial information from internal accounting systems to best estimates of sales managers concerning market potential.

It would be the rare manager using these decision support systems who would possess an explicit or even intuitive knowledge of the quality of the array of data employed by these systems. In response to this need, researchers have been exploring

mechanisms for tagging the content of these databases to indicate the information quality of various elements [Wang and Madnick, 1990].

Once the technical problems of data tagging have been resolved, the important question remains concerning the content of these tags. In keeping with the underlying premise of decision support system design, it is logical that this determination should start with the decision-maker. Consequently our goal is to assist those who must design databases to support decision processes by providing insight into what type of data quality information would have the most constructive impact on decision makers.

Earlier work involving the impact of errors in data on the correctness of decision making was conducted by Ballou and Pazer [1990]. That work, however, was not empirical and did not investigate the interrelationship between the type of decision problem and the most effective presentation of information on data quality.

The currently reported research is the third and final stage of a pilot study exploring the dimensions of a research design developed by the authors in conjunction with Don Ballou. The first step reported in the Proceedings of the 1997 Conference on Information Quality [Chengalur-Smith et al] focused on the impact of data quality tagging on decision complacency. The second stage, currently under review at IEEE/TKDE [Chengalur-Smith et al, 1998], expands the analysis to also explore, in addition to decision complacency, the issues of decision consensus and decision consistency. The current study views the same general construct but also explores the impact of individual differences as measured by differences in learning styles upon decision complacency, consensus and consistency. These pilot studies all employed seniors in our MIS program to fully test the design prior to applying it to practicing managers.

The research design was described in detail on the two above cited papers and will be presented in summary form below. The three factors in the overall research design are decision strategy (conjunctive or weighted additive), task (simple or complex) and data quality information format (no data quality information, two-point ordinal information, or 100-point interval scale information). The current study suppressed the decision strategy dimension (only the conjunctive case was explored) and added the

learning style factor. A more detailed description of the design variant for the current study will be presented later in this paper.

As indicated above, our focus is on the impact of data quality information upon the decision process and more specifically upon the selection of alternative based strategies. Within this context we define three key measures:

Decision Complacency - If the decision-maker does not change the originally preferred alternative after viewing data quality information we say that he/she has exhibited decision complacency. In a sense, decision complacency is a measure of futility for it implies that our effort to provide data quality tags has not impacted the final decision. Obviously low levels of decision complacency are the most desirable outcome. It is important to note that the measure of decision complacency focus only on changes of the top choice.

Decision Consensus - Often a number of decision-makers will be involved in the same decision process. This factor explores the impact of data quality information upon the ability of such groups to maintain the prior degree of agreement concerning the preferred course of action. This is a measure of group response to this new incremental information. Once again, the focus is only on the top choice. The current study explores at the disaggregate level the impact of individual learning style on the response to data quality information.

Decision Consistency - The previous two measures focused only on the top or preferred alternative. There are, of course, numerous decision contexts where the interest is in the ranking of all alternatives, for example, the allocation of merit raises across an entire department. In a sense, consistency is an extension of complacency to the entire set of alternatives. Once again a high value for this measure would indicate that major re-orderings in the ranks did not occur.

Data Quality Metadata

Though it is clear that the metadata describing a database can include information concerning data quality, the detail and content of this subset of metadata is still an open question. Through the use of data tags such as those proposed by Wang and Madnick [1990] data quality information can be provided at the level of the individual data item.

At the other extreme a summary statistic could be provided concerning the quality of an entire file or relational table. An intermediate approach, which we have adopted for this sequence of studies, is based on the provision of a data quality indicator for each of the attributes (data fields) made available to the decision maker. The first approach would be the most complex and costly but may be necessary where the processes generating the data items were erratic and highly volatile. The single indicator approach would be most appropriate where the entire file was generated by a single, rather stable, process. Our approach is most appropriate where the data being utilized originated from a number of processes which may have substantially different data quality capabilities.

The information concerning data quality can be either qualitative or quantitative in nature. Our studies compare the utilization of the highly quantitative approach of describing data quality on a 100-point scale to the semi-qualitative approach of utilizing only the indicators "above average" and "below average". While the granularity of the 100-point scale allows the drawing of fine distinctions between levels of data quality it is an open and researchable question as to whether this level of detail will be effectively utilized by the decision-makers viewing this data.

Obviously the resources required to provide above average vs. below average data quality indicators for a potentially broad array of data generating processes will be substantially less than would be required to scale these processes with finer granularity. Consequently it is important for data base designers to know if this increased investment is justified by positive impact on the decision process.

Rationale for inclusion of Learning Style factor

The type of data quality information that is appropriate may be a function of the task at hand as well as individual preferences. In order to capture and study individual differences in using this kind of information we needed a surrogate for the mindsets of different professionals in an organization. So we added another dimension to this study namely learning style (Kolb et al, 1976). The Learning Style Inventory (LSI) is a simple self-description test that measures a person's relative emphasis on four stages of learning: *concrete experience, reflective observation, abstract conceptualization* and *active experimentation.* Abstract conceptualization (AC) and concrete experience (CE) are

paired to create a scale and active experimentation (AE) and reflective observation (RO) form a second scale. Different combinations of these four learning styles can be characteristic of different professionals in an organization. Some major findings by Kolb et al (1976) are described below:

People who specialize in the physical sciences rely primarily on AC and AE learning styles. For instance, engineers and technical specialists tend to have dominant AC and AE learning modes. Those with humanities or liberal art backgrounds emphasize CE and RO learning approaches. In organizations, personnel managers are characterized by dominant CE and RO learning styles. Individuals in basic science as opposed to applied sciences have learning styles that emphasize AC and RO. In organizations such individuals are often found in research and planning departments. People with an educational background in practical fields such as business are characterized by CE and AE learning styles. In organizations such people are generally found in the marketing or sales departments.

Research Design

This final stage of our pilot study used the same tasks as the previous stage. The apartment selection task had four alternatives with five criteria: parking facilities, commuting time to work, floor space, number of bedrooms, and rent expense. The task was designed to be a relatively simple task with the four choices being clearly distinct. For each alternative the evaluation for each criterion was provided and the minimum acceptable levels for each criterion were given alongside. The other task required subjects to select a site for a restaurant from a set of candidate sites. This task had six alternatives and seven attributes per alternative. Each alternative was evaluated on the following seven criteria: area retail sales, traffic density, competition, average family income, land and building costs, population density, and population growth. The same type of information was provided to the test subjects as with the simple task. However, here the choices were more prone to interpretation, and this task had over twice the number of cells (6 x 7 = 42) as compared to the apartment selection task (4 x 5 = 20).

We again explored the impact of two kinds of information quality formats, namely two-point ordinal and 100-point interval. Within a given task, the data was

identical except for the data quality format and to ensure consistency between formats, we assigned the ordinal designation "Above average" to the highest 50% of the numerical values on the interval scale and "Below average" to the lowest 50%. We were constrained in terms of the number of factors we could explore in this study because of the small sample sizes. Since the previous study resulted in the conjunctive decision making task having more interesting results, we decided to eliminate the weighted additive task from this study.

Subjects were randomly assigned to one of the two tasks. In order to more closely track an individual's thinking process, we initially presented the subjects with a task with no data quality information and then presented the same individual with the same task but with data quality information in one of the two formats. In addition, each subject completed the LSI form. Our subjects were undergraduate students that had completed a database course in the School of Business. They were primarily MIS majors but also included those that were double majoring in marketing and finance.

Hypotheses

The three measures we are interested in are: decision complacency, decision consensus and decision consistency. Recall that complacency (not changing the originally preferred alternative in the presence of data quality information) and consensus (ability to converge on an alternative using data quality) deal with the issue of the top-ranked alternative. The first two sets of hypotheses given below deal with the top ranked choice and the third set of hypotheses deal with the entire set of rankings to test the issue of consistency.

Complacency:

H1o: Including data quality information has no impact on the number of times the originally preferred site continues to be ranked the top site for the simple (complex) task.

H1a: Including data quality information changes the number of times the originally preferred site continues to be ranked the top site for the simple (complex) task

A significant result leading to the rejection of the null hypothesis suggests that the data quality information provided have been utilized.

Consensus:

H2o: Including data quality information has no impact on the number of times the selected site is ranked the top site for the simple (complex) task.

H2a: Including data quality information changes the number of times the selected site is ranked the top site for the simple (complex) task.

The null hypothesis is desirable since it implies that inclusion of data quality information is not detrimental to consensus building.

Consistency:

H3o: No significant correlation exists between the average of the ranks assigned to a site with and without data quality information for the simple (complex) task.

H3a: Significant correlation exists between the average of the ranks assigned to a site with and without data quality information for the simple (complex) task.

The alternative hypothesis suggests that data quality information has not caused major revision in the ordering of the ranks.

Research Results

Tables 1, 2 and 3 present the results of this study for the three measures complacency, consensus and consistency and compare them to the results of last year's study. We find that the results from the two studies are compatible. In terms of complacency the biggest change arose when subjects performed the simple task and were presented with interval information. Thus our subjects were able to utilize data quality information best for this particular scenario. There were no significant differences in consensus for both studies. This is good since it shows that providing data quality information does not destroy consensus between individuals. For consistency the correlation between ranks showed the highest conformance when subjects performed the simple task and were presented with ordinal information. This was true for both studies and showed that there were essentially no switches among average site rankings in the presence of data quality information for this particular scenario.

These aggregate results are then broken down by learning style in order to look for differences among them. Table 4 tracks the top choice for each individual and categorizes individuals as concrete or abstract learners and reflective or active learners. Table 5 does the same for all rankings. Changes are measured by summing the absolute differences between the prior rankings and the revised rankings. The overall rankings are said to have changed only when the sum of absolute deviations is greater than two. The results show that there are no significant differences that are attributable to learning style. This makes for simpler implementation since this implies that the database does not have to be tailored to individual departments.

Implications

In the 1997 study we reported: "Overall, we find that in a situation where decision makers are confronted with clearly differentiated alternatives, that the inclusion of data quality information impacted the selection of a preferred alternative while maintaining group consensus. The format of presentation of this data quality information seems important with an indication that when a more complex format is used in a more complex environment that information overload may occur." [Chengalur-Smith et al, 1998]

While part of the support for the above statement came from the analysis of the weighted additive decision strategy which was not replicated on this study, nothing in the results of the current study have contradicted these findings. Once again we find that a greater impact of data quality information is observed in the simpler task. In fact, for the current study this is the only situation in which a significant rejection of decision complacency occurred. The earlier study was strongly suggestive of the existence of information overload and it once again is a possible explanation for our observed results.

In both studies when the aggregate ranking of the various alternatives are compared with and without data quality information, a high correlation is observed. While it was relatively common for the inclusion of data quality information to switch the rankings of adjacent pairs, it was rare for changes greater than this to occur. This causes us to reaffirm the following statement made in our analysis of the 1997 study.

"This seems to suggest that the inclusion of data quality information is more valuable in situations where the decision maker needs to find the best alternative while it may be less

valuable when the decision process calls for the comparison of averages of larger subgroups which would be impacted only at the margins." [Chengalur-Smith et al, 1998]

In no case was the impact of learning style significant at even the 0.05 level. In fact the largest chi-square value which corresponded to the concrete/abstract comparison for the restaurant (complex) scenario with interval data quality information was only 2.03. This would not even have been significant at a 0.10 level. To the extent that learning style as measured by the LSI serves as a surrogate for individual differences between professionals in an organization, this is a reassuring result for the database designer. If this finding is reaffirmed by a follow-up study using an array of professionals in actual organizations then it indicates that at least in terms of data quality information that "one size may fit all". This is important since the same database may be used by decision-makers in a number of functional areas of the organization.

Research design for future studies

The format of presentation of data quality information seems important with an indication that when a more complex format is used in a more complex environment that information overload may occur. It is, however, important to recall that subjects of this experiment were seniors in an undergraduate MIS program. Information overload may be a relative concept. What is considered too complex and information intensive to undergraduates may seem less daunting to analysts and decision-makers accustomed to real world complexities. This is one of the issues that should be clarified in the next stage of our research.

We are currently seeking a business or government agency to serve as the site for the next stage of this research. Once the site is identified, we will develop a simple and a complex task closely replicating decision scenarios encountered in this environment. These decision tasks will utilize a combination of financial and non-financial information. These will be pretested using seniors in our MIS program.

Instead of relying on the LSI as a surrogate we will utilize three subgroups of employees such as from finance, marketing and MIS. For half of these individuals we will test the two point ordinal format for presenting data quality information while for the

rest we will test the more detailed 100-point scale. The results of this extended study may provide important additional insights to database designers.

References

(1) Ballou, D. P. and H. L. Pazer. "Framework for the analysis of error in conjunctive, multi-criteria, satisficing decision processes" *Decision Sciences* 21,4 (1990), 752-770.

(2) Chengalur-Smith, I. N., D. P Ballou and H. L. Pazer. "The impact of data quality tagging on decision complacency" in *Proceedings of the 1997 Conference on Information Quality*, Cambridge, MA, pp. 209-221.

(3) Chengalur-Smith, I. N., D. P Ballou and H. L. Pazer. "The impact of data quality information on decision making: An exploratory analysis" under review at *IEEE Transactions on Knowledge and Data Engineering* (1998).

(4) Kolb, D., Rubin, I., and McIntyre, J. (eds) *Organizational Psychology: A Book of Readings*, 2nd ed., Englewood Cliffs, NJ Prentice Hall, 1974.

(5) Wang, R. Y. and S. E. Madnick. "A polygon model for heterogeneous database systems: The source tagging perspective." In *Proceedings of the 16th International Conference on Very Large Databases*, (1990) Brisbane, Australia, pp. 519-538.

Table 1: Analysis of Complacency

Task	*Data Quality Information*	*1997*	*1998*
Simple (4 alternatives x 5 attributes)	None versus Ordinal	Not significant	Not significant
	None versus Interval	Significant	Significant
Complex (6 alternatives x 7 attributes)	None versus Ordinal	Marginally significant	Not significant
	None versus Interval	Not significant	Not significant

Table 2: Analysis of Consensus

Task	*Data Quality Information*	*1997*	*1998*
Simple (4 alternatives x 5 attributes)	None versus Ordinal	Not significant	Not significant
	None versus Interval	Not significant	Not significant
Complex (6 alternatives x 7 attributes)	None versus Ordinal	Not significant	Not significant
	None versus Interval	Not significant	Not significant

Table 3: Analysis of Consistency

Task	*Data Quality Information*	*1997*	*1998*
Simple (4 alternatives x 5 attributes)	None versus Ordinal	CORR = 0.99	CORR = 0.99
	None versus Interval	CORR = 0.91	CORR = 0.92
Complex (6 alternatives x 7 attributes)	None versus Ordinal	CORR = 0.86	CORR = 0.95
	None versus Interval	CORR = 0.96	CORR = 0.96

Table 4: Analysis of top ranked sites categorized by LSI

Restaurant

Interval

	No movement	Movement
concrete	9	3
abstract	4	5

	No movement	Movement
reflective	7	5
active	6	3

Ordinal

	No movement	Movement
concrete	2	9
abstract	3	4

	No movement	Movement
reflective	2	9
active	3	4

Apartment

Interval

	No movement	Movement
concrete	5	4
abstract	3	3

	No movement	Movement
reflective	4	5
active	4	2

Ordinal

	No movement	Movement
concrete	6	6
abstract	2	3

	No movement	Movement
reflective	4	7
active	4	2

Cell sizes may be reduced due to missing values

Table 5: Analysis of all ranks categorized by LSI

	Restaurant	
Interval		
	No movement	Movement
concrete	5	7
abstract	2	7
	No movement	Movement
reflective	4	8
active	3	6
Ordinal		
	No movement	Movement
concrete	5	6
abstract	2	5
	No movement	Movement
reflective	3	8
active	4	3

	Apartment	
Interval		
	No movement	Movement
concrete	5	4
abstract	3	3
	No movement	Movement
reflective	5	4
active	3	3
Ordinal		
	No movement	Movement
concrete	10	2
abstract	4	1
	No movement	Movement
reflective	9	2
active	5	1

Cell sizes may be reduced due to missing values