

## **Using Statistical Metadata to Adjust for Variances and Systematic Biases: An Exploratory Study**

A. Faye Borthick, DBA, CISA, CMA, CPA  
School of Accountancy  
Georgia State University  
Atlanta GA 30302-4050  
Voice: 404 651-4472; Fax: 404 651-1033  
[borthick@gsu.edu](mailto:borthick@gsu.edu)

Paul L. Bowen, PhD, CPA  
College of Business  
Florida State University  
Tallahassee FL 32306-1110  
Voice: 850 644-4224; Fax: 850 644-8234  
[pbowen@cob.fsu.edu](mailto:pbowen@cob.fsu.edu)

Gregory J. Gerard, PhD  
College of Business  
Florida State University  
Tallahassee FL 32306-1110  
Voice: 850 644-9115; Fax: 850 644-8234  
[ggerard@cob.fsu.edu](mailto:ggerard@cob.fsu.edu)

Jon D. Perkins, JD, PhD, CMA, CPA  
College of Business  
University of Illinois  
Champaign, IL 61820  
Voice: 217 333-4527; Fax: 217 333-0902  
[jdperkin@illinois.edu](mailto:jdperkin@illinois.edu)

David A. Robb, PhD  
UQ Business School  
University of Queensland  
Brisbane, Queensland, Australia 4072  
Voice: +61-7-3381-1219; Fax: +61-7-3365-7285  
[a.robb@business.uq.edu.au](mailto:a.robb@business.uq.edu.au)

1 August 2008

# Using Statistical Metadata to Adjust for Variances and Systematic Biases: An Exploratory Study

## Abstract

Current advances in technology, e.g., data marts and data warehouses, have increased the amount of information available to decision makers. Coping with the explosion of data has put an increased emphasis on the use of metadata (data about data) to help individuals incorporate this information into their decision-making processes. Prior research on metadata has focused on the underlying data, not on types and characteristics of the metadata itself. We investigate the impact of providing various types of metadata on individuals' judgments. We also explore characteristics of the decision-making environment such as decision frame. Preliminary results support our contention that individuals are not good at incorporating dispersion metadata (e.g., range and standard deviation information) into their decision-making processes, but that they are better able to incorporate standard deviation data than range data. Our preliminary results also support our belief that the type of decision frame (gain or loss) interacts with the type of metadata provided to affect individuals' performance.

(Metadata, Information Quality, Data Quality, Data Quality Information, Bias, Statistical-Based Adjustments; Decision Making)

# Using Statistical Metadata to Adjust for Variances and Systematic Biases: An Exploratory Study

## Introduction

Over the years, technology has allowed individuals to access more data to assist them in their decision-making processes. When faced with this explosion of available data, individuals may use metadata (data about data) to evaluate the quality of the data and make adjustments that produce higher quality decisions. Common metadata tags can include systematic biases and dispersion metrics such as the typical range and standard deviation of various elements of the data. How effectively individuals use this statistical metadata depends on the type of metadata provided, the decision context, and the decision frame.

Prior research investigating the effects of metadata on decision quality has focused on metadata about the quality of the data, not metadata about the data itself. To our knowledge, this is the first study to explore the effects of range and standard deviation metadata tags on individuals' judgments as well as the first study to examine the interactive effects of these tags with decision frame and decision context. We also contribute to the literature through our investigation of how individuals use these types of metadata information in their decision-making processes.

## Background

### *Statistical Metadata*

When managers make decisions, they often have access to a large amount of data they can use to assist them. To aid them in making effective decisions, they are often

given metadata in the form of summary statistical information. One common piece of metadata is information about the range of the underlying information. Prior research has shown that individuals provided with range estimates typically use the mid-point of the range as a point estimate. Unfortunately, we found no research showing what type of distribution, e.g., uniform, normal, or exponential, these individuals tend to use.

Providing individuals with standard deviation information in addition to range information should allow us to begin investigating the type and shape of the distributions decision makers assume. If individuals initially assume a uniform distribution, providing standard deviation metadata may cause them to change their assumed distribution, e.g., to an assumption of a normal distribution, which would lead to a change in the decision makers judgments. This change in the assumed distribution may occur even though provision of standard deviation information should not necessarily change individuals' beliefs about the underlying information distribution. This change in the assumption about the distribution of the underlying data should cause individuals to become more confident in their decisions, i.e., because the newly-assumed distribution has a more precisely defined shape.

### *Prior Experiments*

Chengalur-Smith et al. (1999) conducted an experiment regarding the consequences of providing metadata about the quality of data available for decision making tasks. Two tasks were used: a relatively simple apartment selection task and a more complex restaurant task. Chengalur-Smith et al. used these experimental tasks to explore decision complacency, decision consensus, and decision consistency. They

found that participants (students) assigned to the simple task made greater use of data quality metadata (i.e., complex task participants were more complacent). Minimal effects for data quality meta data were observed relative to consensus or consistency (i.e., data quality meta data had little effect on simple versus complex tasks or for the presence or absence of data quality metadata) Overall, data quality metadata differences were associated with changes in complacency but few statistical differences relative to consensus or consistency. Of particular note, Chengalur-Smith et al. (page 681) assert that “Providing too detailed information concerning data quality may be counterproductive in more complex decision environments.” Chengalur-Smith et al. also note conjecture that more experience may mitigate complexity, i.e., information overload is relative to expertise.

Fisher et al. (2003) also conducted two experiments that addressed the impacts of data quality information relative to complacency, consensus, and consistency. Their experiments focused on the effects of differences in experience and time pressure. Two tasks were used: a simple task that replicated the simple (20 cells) Chengalur-Smith apartment select task and a new, more complex job transfer task. The job transfer task was more complex than Chengalur-Smith’s restaurant site selection task (42 cells for the restaurant site selection task versus the 63 cells associated with the new job transfer task). The first task used both novices and experts and the second task only used experts. Results of the first (simple) task indicated no significant differences between novices and experts in the no data quality information (DQI) condition. When DQI was available to participants, experienced participants made substantial use of the DQI but the novices did

not. Results of the second (complex) task revealed that participants with greater management experience made substantially greater use of the DQI.

The Chengalur-Smith et al. (1999) and the Fisher et al. (2003) experiments primarily investigated whether or not particular participant groups exhibited complacency, consensus, and consistency in their use of DQI, e.g., these experiments take a somewhat global view of participants' use of DQI. The research reported in this paper explores whether particular kinds of DQI, i.e., statistical metadata, are or are not useful to particular participant groups. Furthermore, relative to the Chengalur-Smith et al. (1999) and the Fisher et al. (2003) experiments, the experimental tasks reported in this research are more focused on recurring business forecast and production (operating) decisions rather than more personal or strategic decisions (apartment selection, restaurant site selection, and job transfer).

### *Decision Context*

At least two prior studies have found that dispersion information has had no significant effect on decision making (Oliver 1972; Birnberg and Slevin 1976). One explanation for their results is that, in the decision context they used, the users were unable to incorporate the dispersion information into their mental models of that problem. We will investigate if decision context interacts with the provision of metadata information to affect individuals' judgments. In a business decision context such as a production decision, individuals are more likely to be able to incorporate standard deviation information into their mental models than in a non-business decision context such as purchasing a car, choosing an apartment, or similar decision that includes many

personal preference criteria. Further, our investigation of a business setting will add greater external validity to prior metadata studies that have generally used non-business contexts.

### *Decision Frame*

Prospect theory (Kahneman and Tversky 1979) suggests that the framing of a decision as a gain or a loss affects how individuals code possible outcomes. In a similar way, such framing may also affect if and how individuals use metadata information. As a result, we will investigate if framing the decision to be made as a gain or loss interacts with the provision of metadata information to affect individuals' judgments.

### *Individuals' Use of Metadata Information*

To our knowledge, no prior study has examined the process by which individuals use metadata information in their judgments. Not only will we investigate the effects of the metadata information on judgments, but we will also delve into the "black box" and investigate how the information is used to arrive at those judgments.

## **Hypothesis Development**

### *Type of Metadata Provided*

Depending on their ability to incorporate dispersion information into their decision-making processes, providing individuals with information about the dispersion of the underlying data should improve their judgments,. Making a judgment using dispersion information is inherently complex and difficult. For example, compared to a

single point estimate for a mean (or median), a range not only focuses on two end points but raises substantial uncertainty about the density of the population between those two points. Therefore, we predict that individuals provided with dispersion information (e.g., range and/or standard deviation) information will make judgments of poorer quality than individuals who are not provided that information.

H1: When confronted with decisions with asymmetric profit/loss functions, compared with decision makers with point estimate and bias information, decision makers with range and bias information will (a) exhibit greater variation in their recommendations (lower consensus), (b) make less optimal adjustments, and (c) be less confident that they recommended an optimal adjustment.

H2: When confronted with decisions with asymmetric profit/loss functions, compared with decision makers with point estimate and bias information, decision makers with point estimate, bias, and standard deviation information will (a) exhibit greater variation in their recommendations (lower consensus), (b) make less optimal adjustments, and (c) be less confident that they recommended an optimal adjustment.

H3: When confronted with decisions with asymmetric profit/loss functions, compared with decision makers with point estimate and bias information, decision makers with range, bias, and standard deviation information will (a) exhibit greater variation in their recommendations (lower consensus), (b) make less optimal adjustments, and (c) be less confident that they recommended an optimal adjustment.

Providing individuals with information about the standard deviation of the underlying information will have a different effect on judgments than providing them with range information. Prior research has shown that individuals given range information tend to assume a uniform distribution. Further, we believe that individuals given standard deviation information are likely to assume a normal distribution because standard deviation information is more closely associated with normally distributed data. Because normally distributed data is more clustered around a point estimate, we believe that individuals provided with standard deviation, point estimate, and bias information will make better judgments than individuals provided with range and bias information.

H4: When confronted with decisions with asymmetric profit/loss functions, compared with decision makers with range and bias information, decision makers with point estimate, bias, and standard deviation information will (a) exhibit less variation in their recommendations (lower consensus), (b) make more optimal adjustments, and (c) be more confident that they recommended an optimal adjustment.

The effect of providing both range and standard deviation information in addition to point estimate and bias information is unclear. The additional information should result in improved judgments, but individuals may have difficulty incorporating these seemingly conflicting distributional assumptions (one suggesting a normal distribution and the other suggesting a uniform distribution) into their decision-making processes.

Accordingly, we do not predict any difference in judgment quality when providing both range and standard deviation information.

H5: When confronted with decisions with asymmetric profit/loss functions, compared with decision makers with range and bias information, decision makers with range, bias, and standard deviation information will (a) exhibit similar variations in their recommendations (lower consensus), (b) make similar adjustments, and (c) be similar in their confidence that they recommended an optimal adjustment.

H6: When confronted with decisions with asymmetric profit/loss functions, compared with decision makers with point estimate, bias, and standard deviation information, decision makers with range, bias, and standard deviation information will (a) exhibit the similar variations in their recommendations (lower consensus), (b) make similar adjustments, and (c) be similar in their confidence that they recommended an optimal adjustment.

### *Form of Bias*

Prior research has shown that individuals' judgments are affected by their decision frame. One context that is common in the provision of metadata is whether the data are biased in one direction or another. For example, in making production decisions, data could be biased towards overproduction or underproduction. In a normal business setting with an asymmetrical profit/loss function, overproduction bias results in a loss frame (causing individuals to become more risk-seeking in their judgments) while

underproduction bias results in a gain frame (causing individuals to become less risk-seeking in their judgments). Therefore, an overproduction bias should have more of an effect on individuals' judgments than an underproduction bias (i.e., judgments in an overproduction bias setting should be of lower quality than judgments in an underproduction bias setting).

H7: When confronted with decisions with asymmetric profit/loss functions, compared with decision makers with underproduction bias information, decision makers with overproduction bias information will (a) exhibit greater variation in their recommendations (lower consensus), (b) make less optimal adjustments, and (c) be less confident that they recommended an optimal adjustment.

## **Experiment**

To investigate the above hypotheses, we have run an experiment where 107 undergraduate student participants from two major universities each completed four short cases (one for each of the four metadata conditions). Each participant was in either the Overproduction Bias condition or the Underproduction Bias condition (i.e., the metadata information provided was manipulated within subjects while the form of bias was manipulated between subjects). Although the students were not given a time limit for completion of the cases, most students completed the four cases in approximately 30-40 minutes. The experimental task is discussed in more detail below.

### *Experimental Task*

The following task can be adjusted for any seasonal situation, e.g., Christmas trees, Halloween candy, or Mother's day cut flowers, and for any proportion of profits and losses as long as the profit per unit and the loss per unit are not equal. That is, the decision is obvious if the profit per unit and the loss per unit are equal.

#### Rudolph's Christmas Tree Farm

Treatment: Control (point estimate and bias)

Each year, Rudolph's Christmas Tree Farm (located in Canada) cuts and ships Christmas trees to various customers such as Lowe's, Home Depot, Wal-Mart, and Sears. RCTF makes a profit of \$14 on each tree shipped and sold. Due to labor, shipping, and disposal costs, each tree shipped but not sold incurs a \$6 loss. RCTF's management believes customers consistently order an average of 100,000 less than they could sell. Total orders this year are for 1,000,000 trees. How many trees should RCTF cut and ship this year?

Treatment: Range and bias

Each year, Rudolph's Christmas Tree Farm (located in Canada) cuts and ships Christmas trees to various customers such as Lowe's, Home Depot, Wal-Mart, and Sears. RCTF makes a profit of \$14 on each tree shipped and sold. Due to labor, shipping, and disposal costs, each tree shipped but not sold incurs a \$6 loss. Sales estimates for this year's sales range from 900,000 to 1,100,000 trees. RCTF believes these estimates are 100,000 too low. How many trees should RCTF cut and ship this year?

Treatment: Point estimate, bias, and standard deviation

Each year, Rudolph's Christmas Tree Farm (located in Canada) cuts and ships Christmas trees to various customers such as Lowe's, Home Depot, Wal-Mart, and Sears. RCTF makes a profit of \$14 on each tree shipped and sold. Due to labor, shipping, and disposal costs, each tree shipped but not sold incurs a \$6 loss. RCTF's management believes customers consistently order an average (standard deviation) of 100,000 (50,000) less than they could sell. Total orders this year are for 1,000,000 trees. How many trees should RCTF cut and ship this year?

Treatment: Range, bias, and standard deviation

Each year, Rudolph's Christmas Tree Farm (located in Canada) cuts and ships Christmas trees to various customers such as Lowe's, Home Depot, Wal-Mart, and Sears. RCTF makes a profit of \$14 on each tree shipped and sold. Due to labor, shipping, and disposal costs, each tree shipped but not sold incurs a \$6 loss. Sales estimates for this year's sales range from 900,000 to 1,100,000 trees. RCTF believes these estimates are an average (standard deviation) 100,000 (50,000) too low. How many trees should RCTF cut and ship this year?

Treatments

1. Control: Point estimate and bias (PB)
2. Range and bias (RB)
3. Point estimate, bias, and standard deviation (PBS)
4. Range, bias, and standard deviation (RBS)

Model:

Forecast accuracy, confidence = information set {PB, RB, PBS, RBS} + context  
 {personal, corporate} \* information set + frame {gain/loss} + covariates

Covariates = {GPA, business experience, statistical training, gender}

## Data Analysis

### *Descriptive Statistics*

Below are the descriptive statistics. The first number in each cell is the mean of the scaled scores for that cell (explained further below), the number in parentheses is the standard deviation of those scores, and the number in brackets is the number of participants in that cell.

| Form of Bias    | Data Provided                   |                                |                                 |                                 |
|-----------------|---------------------------------|--------------------------------|---------------------------------|---------------------------------|
|                 | PE & Bias                       | Range & Bias                   | SD, PE, & Bias                  | Range, SD, & Bias               |
| Overproduction  | 0.500686<br>(0.841432)<br>[52]  | 1.379162<br>(1.348145)<br>[49] | 0.895453<br>(0.641526)<br>[54]  | 1.305314<br>(1.098539)<br>[54]  |
| Underproduction | 0.484875<br>(0.824542)<br>[51]  | 1.082531<br>(1.031004)<br>[48] | 0.713118<br>(0.465856)<br>[53]  | 0.880055<br>(0.736166)<br>[53]  |
|                 | 0.492857<br>(0.829057)<br>[103] | 1.232375<br>(1.204734)<br>[97] | 0.805138<br>(0.566243)<br>[107] | 1.094671<br>(0.956496)<br>[107] |

Because the cases were different across experimental conditions (the numbers used in each case were different), we scale participants' responses for each case to make them comparable. To do this, we took each participant's raw response, subtracted the average of the sales estimates as well as the bias adjustment, divided that result by the bias adjustment, and took the absolute value of that amount:

$$\text{Scaled score} = \left| \frac{(\text{Participant Response} - \text{Sales Average}) - \text{Bias Adjustment}}{\text{Bias Adjustment}} \right|$$

So, for example, in the PE & Bias with Overproduction Bias cell above, the average of the sales estimates was 1,000,000 units and the overproduction bias was 20,000. If a participant's response was to produce 950,000 units, the scaled score would be calculated as follows:

$$\text{Scaled score} = \left| \frac{(950,000 - 1,000,000) - (-20,000)}{(-20,000)} \right| = 1.5$$

The basic interpretation of the scaled score is that the closer that a participant's response is to the optimal answer, the closer the scaled score is to zero.

### *Preliminary Data Analyses*

To facilitate description of the preliminary data analyses, we reproduce our experimental design below:

| Form of Bias    | Data Provided |              |                |                   |
|-----------------|---------------|--------------|----------------|-------------------|
|                 | PE & Bias     | Range & Bias | SD, PE, & Bias | Range, SD, & Bias |
| Overproduction  | A             | D            | G              | J                 |
| Underproduction | B             | E            | H              | K                 |
|                 | C             | F            | I              | L                 |

To investigate the main effect of the type of data provided to participants, we ran basic t-tests to compare the PE & Bias condition (our "control" condition) to each of the other three conditions:

| Hypothesis | Comparison | t-statistic | p-value (one-tailed) | p-value (two-tailed) |
|------------|------------|-------------|----------------------|----------------------|
| H1         | C & F      | 5.02761     | < 0.001              | < 0.001              |
| H2         | C & I      | 3.17569     | 0.001                | 0.002                |

|    |       |          |         |         |
|----|-------|----------|---------|---------|
| H3 | C & L | 4.87759  | < 0.001 | < 0.001 |
| H4 | F & I | 3.188051 | 0.001   | 0.001   |
| H5 | F & L | 0.898033 | 0.185   | 0.370   |
| H6 | I & L | 2.69443  | 0.004   | 0.008   |

As a result, it appears that participants' performance did decrease as the amount of information provided increased. The higher scores for the cells that included the range data lend some support to our thought that when provided with range information, the participants interpreted the underlying data as being uniformly distributed, which would make the optimal answer more difficult to calculate than if they assumed a normal distribution. The significance for the F & I comparison as well as the I & L comparison also lend support to that conclusion (even though we did not expect any results for H6).

While we have not yet run an ANOVA to investigate the main effect for the form of bias, we ran basic t-tests as follows to investigate the effect of form of bias on each of the data conditions:

| Comparison | t-statistic | p-value (one-tailed) | p-value (two-tailed) |
|------------|-------------|----------------------|----------------------|
| A & B      | 0.09632     | 0.462                | 0.923                |
| D & E      | 1.21877     | 0.113                | 0.226                |
| G & H      | 1.68453     | 0.048                | 0.095                |
| J & K      | 2.35626     | 0.010                | 0.020                |

As a result, it appears that there is some interaction between the form of the bias and the data provided. As more data is provided, the overproduction bias appears to have a greater effect on participants' performance.

We plan to supplement these preliminary analyses with some ANOVA analyses to investigate these comparisons further as well as to investigate the effects of other variables such as the school that the participants attend. We also will investigate the

effect of these items on individuals' confidence in their judgments as well as their consensus. Further, we plan to expand our experiment to investigate the effect of context (business versus non-business) on individuals' use of metadata information.

### **Planned Conference Presentation**

In our presentation, we plan to present our motivation for our study and the contribution that it makes to the current body of literature on the subject. We also plan to present the basic theory for our hypotheses as well as discuss our experimental instruments and preliminary results. We plan to wrap up our presentation with a discussion of the unfinished portion of the project, including any additional planned experiments and additional statistical analyses.

### **References**

- Birnberg, J.G. and D.P. Slevin. 1976. A note on the use of confidence interval statements in financial reporting. *Journal of Accounting Research* 14 (Spring): 153-157.
- Chengalur-Smith, I.N., D.P. Ballou, and H.L. Pazer. 1999. The impact of data quality information on decision making: An exploratory analysis. *IEEE Transactions on Knowledge and Data Engineering* 11, (6) (November-December): 853-864.
- Fisher, C.W., I. Chengalur-Smith, and D.P. Ballou. 2003. The impact of experience and time on the use of data quality information in decision making. *Information Systems Research* 14, (2) (June): 170-188.
- Kahneman, D. and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47 (March): 263-292.

Oliver, B.L. 1972. A study of confidence interval financial statements. *Journal of Accounting Research* 10 (Spring): 154-166.

## Appendix Analytical Solution for Proposed Experimental Task

Variables

M = units manufactured (decision variable, equivalent to choice of p)

PPU = profit per unit if the unit is sold

LPU = loss per unit if the unit is not sold

MR = maximum range, i.e., the maximum number of units possible to be sold

p = proportion (percent) of the maximum number of units possible to be sold that are manufactured (decision variable, equivalent to choice of M)

1-p = proportion (percent) of the maximum number of units possible to be sold that are not manufactured

$E(\pi_M)$  = expected profit of manufacturing M units

{? Jon, can we include one or more figures?}

$$E(\pi_M) = M \times [(p/2 \times PPU) - (p/2 \times LPU) + ((1-p) \times PPU)]$$

Because  $M = p \times \text{Maximum Range}$ ,

$$\begin{aligned} E(\pi_M) &= p \times MR \times [(p/2 \times PPU) - (p/2 \times LPU) + ((1-p) \times PPU)] \\ &= (p \times MR)/2 \times [(p \times PPU) - (p \times LPU) + (2 \times PPU) - (2 \times p \times PPU)] \\ &= (p \times MR)/2 \times [(2 \times PPU) - (p \times PPU) - (p \times LPU)] \\ &= (p \times MR \times PPU) - (p^2 \times MR \times PPU)/2 - (p^2 \times MR \times LPU)/2 \end{aligned}$$

Differentiate by p and set equal to zero to find the maximum value:

$$\partial E(\pi_M) / \partial p = (MR \times PPU) - (p \times MR \times PPU) - (p \times MR \times LPU) = 0$$

$$(MR \times PPU) = (p \times MR \times PPU) + (p \times MR \times LPU)$$

$$PPU = (p \times PPU) + (p \times LPU)$$

$$PPU = p \times (PPU + LPU)$$

$$p = PPU / (PPU + LPU)$$

Example:

Assume

$$PPU = 70$$

$$LPU = 30$$

$$MR = 200$$

Then

$$p = \text{PPU}/(\text{PPU}+\text{LPU}) = 70/100 = 0.7$$

$$M = p \times MR = 0.7 * 200 = 140$$

$$E(\pi_M) = M \times [(p/2 \times \text{PPU}) - (p/2 \times \text{LPU}) + ((1-p) \times \text{PPU})]$$

For  $M = 140$

$$E(\pi_M) = 140 \times [(0.7/2 \times 70) - (0.7/2 \times 30) + (0.3) \times 30]$$

$$= 140 \times [24.5 - 10.5 + 9]$$

$$= 140 \times 23$$

$$= 3220$$