

Determining the use of Data Quality Metadata (DQM) for decision making purposes and its impact on decision outcomes - An Exploratory Study

Helen-Tadesse Moges^a, Véronique Van Vlasselaer^a, Wilfried Lemahieu^a, Bart Baesens^{a,b,c,*}

^aDepartment of Decision Sciences and Information Management, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

^bSchool of Management, University of Southampton, Southampton, SO17 1BJ, United Kingdom

^cVlerick Business School, Leuven, Belgium

Abstract

Decision making processes and their outcomes can be affected by a number of factors. Among them, the quality of the data is critical. Poor quality data causes poor decisions. Although this fact is widely known, data quality (DQ) is still a critical issue in organizations because of the huge data volumes available in their systems. Therefore, literature suggests that communicating the DQ level of a specific data set to decision makers in the form of DQ metadata (DQM) is essential. However, the presence of DQM may overload or demand cognitive resources beyond decision makers' capacities, which can adversely impact the decision outcomes. To address this issue, we have conducted an experiment to explore the impact of DQM on decision outcomes, to identify different groups of decision makers who benefit from DQM and to explore different factors which enhance or otherwise hinder the use of DQM. Findings of a statistical analysis suggest that the use of DQM can be enhanced by data quality training or education. Decision makers with a certain level of data quality awareness used DQM more to solve a decision task than those with no data quality awareness. Moreover, those with data quality awareness reached a higher decision accuracy. However, the efficiency of decision makers suffers when DQM is used. Our suggestion would be that DQM can have a positive impact on decision outcomes if it is associated with some characteristics of decision makers, such as a high data quality knowledge. However, the results do not confirm that DQM should be included in datawarehouses as a general business practice, instead organizations should first investigate the use and impact of DQM in their setting before maintaining DQM in datawarehouses.

Keywords: Data quality, decision strategy, decision support systems, Data Quality Metadata (DQM)

1. Introduction

Although the importance of DQ has been recognized for more than decades, different DQ problems continue to exist even in simple traditional systems because of huge data volumes and their complexity [4]. The problem is exacerbated by the fact that decision support systems are becoming vital to support decision making processes. The DQ level in decision support systems may not be good for different reasons. One reason is that DQ problems can be aggravated when data are merged or integrated from different sources which is typically

the case in decision support systems or datawarehouses. The other reason can be that soft data analysis is needed for strategic planning. Soft data is a subjective assessment or a future trend forecast which can be used for decision making [9]. For example, decision makers need to utilize soft data, such as the marketing strategies of competitors in order to change or adapt the marketing strategy of the company accordingly. Most of the time, managers make decisions without considering the DQ level of the data. Decision makers who are familiar with the data have an intuitive knowledge about the data. However, this intuitive knowledge can be missed when data are used by different decision makers for purposes other than the original purpose for which the data were created, which is becoming more and more the case with the increasing use of datawarehouses. Decision makers who do not have prior experience with the data may avoid using them because they can't verify the quality of

*Corresponding author. Tel. +32 16 32 68 84; Fax +32 16 32 66 24

Email addresses: Helen.Moges@kuleuven.be (Helen-Tadesse Moges), Veronique.VanVlasselaer@kuleuven.be (Véronique Van Vlasselaer), Wilfried.Lemahieu@kuleuven.be (Wilfried Lemahieu), Bart.Baesens@kuleuven.be (Bart Baesens)

the data [9]. Because of such and other reasons, DQ is very important for decision making processes, but organizational datawarehouses are still facing different DQ problems [9].

As one of different ideas to reduce the impact of poor DQ on decision outcomes, the literature suggests the inclusion of metadata about the quality of data (DQM) for two reasons [4, 9, 33, 41]. First, decision makers must adjust their decision making processes accordingly by recognizing the DQ level of the given data [38]. Second, DQ is context-dependent, meaning that data with good quality for one use may not be appropriate for other uses. For instance, the extent to which data is required to be complete for accounting tasks may not be required for sales prediction tasks. Therefore, DQM can help decision makers to determine the appropriateness of the DQ level in the context of the task at hand [25]. Additionally, DQ practitioners have acknowledged the importance of providing DQM to facilitate the decision making process [7, 22].

Maintaining DQM into databases means maintaining the level of DQ measured along DQ dimensions such as accuracy, completeness and timeliness. However, the advantage of providing DQM to decision makers along with the actual data should be fully studied because it would be expensive to collect, maintain and manipulate DQM. Additionally, DQM can be difficult to capture and measure, and may require training and software tools. Moreover, the impact of DQM on decision outcomes can be negative. In response, prior DQM research investigated the use of DQM for decision making processes, although there is no full consensus on the results [4, 9, 23]. Some researchers have found that DQM is used in certain situations [9], and others didn't find any statistical evidence that DQM is actually used, even when it is available [34]. The difference in the results may be attributed to the different approaches used by prior researchers. In addition, the impact of DQM on the effectiveness of decision outcomes is not studied adequately. To fill this gap, this paper investigates the impact of DQM on decision outcomes in a different setting from previous research. We have developed a critical decision task (bankruptcy prediction) based on an Altman-Z model [1] to understand the impact of DQM on the effectiveness of decision outcomes, to identify different groups of decision makers who benefit from DQM and to explore different factors which enhance or otherwise hinder the use of DQM. This study aims to provide a concise set of guidelines for system designers to determine the importance of DQM for their specific case and to justify the associated cost of capturing and maintaining it. The study incorporated all the variables

studied in previous DQM research in addition to novel variables such as DQA which makes the study inclusive. This, in turn, helped to measure the effect of the variables on the use of DQM in a similar environment where similar subjects are used, consequently removing the impact of an experimental design. The main contribution of this study, apart from the inclusion of the DQA variable, is the way in which the decision outcome measures were defined.

The paper is structured as follows. The next section reviews previous research in DQM. The third section discusses the research design and the fourth section explains the results. Finally, the paper ends by giving concluding remarks and indicating future research ideas.

2. Literature Review

2.1. Data Quality

Recently, data quality (DQ) is becoming a concern to organizations where plenty of data are available. Similarly, DQ is constantly growing as a crucial research topic in academic world. DQ research can be categorized into two broad types, *intrinsic* and *contextual* DQ studies. The *intrinsic* DQ research concerns about the intrinsic value of the data. It depends on the data themselves without considering the context in which the data is used. The *contextual* DQ study considers factors such as the purposes for which the data is used and the characteristics of the data users. Prior research has indicated that these contextual factors can strongly affect the way DQ is assessed for daily use. For example, Wang and Strong [46] have indicated the importance of recognizing the multi-dimensionality nature of DQ and measure data items accordingly using users' perceptions. However, the importance of considering contextual DQ assessment may increase the complexity level of DQ management. For example, consider a production company sales sheet which shows "item codes", "quantities", "cost" and "selling prices" where some values for the "cost" column are missing. For decisions regarding production efficiency, the sheet with missing "cost" data would be considered incomplete. However, the same sales sheet can be considered as complete for making inventory decisions (reconciling the amount of quantities on the sheet and the physical quantities in a store) because all the values for the "quantities" column are present. Although not easy, considering the contextual nature of DQ can improve DQ management in databases. In line with this, it is important that decision makers can determine the level of DQ for the task at

hand. This is also one of the reasons why recent DQ research has suggested the integration of DQM along with the data in decision support systems [34].

2.2. Data quality metadata (DQM)

Data quality metadata (DQM) is information about the quality level of stored data in organization databases, and is measured along different dimensions such as accuracy, currency, and completeness. Also, DQM is considered to be intrinsic to the data because the metadata is usually produced objectively. DQ tagging is the process by which DQM is created [40]. There are different types of metadata in information systems which are maintained and managed, such as data dictionary metadata, administrative metadata, and metadata about the system infrastructure (see Table 1).

Types of metadata	Description
Data quality metadata	It indicates the quality level of specific data in databases. For example, it can be indicated that sales data are 90% complete for the month of January 2014.
Descriptive metadata	It describes the data in terms of e.g. purpose, author, title etc.
Terms and conditions metadata	It describes the conditions under which the data can (not) be used, e.g. intellectual property rights.
Administrative metadata	It indicates when and how the data are created, and who can access them.
Data dictionary metadata	It indicates the meaning of and relationships within the data.
Structural metadata	It describes the syntactical aspects of the data, e.g. the structure and base type of the data records.

Table 1: Different types of metadata as discussed in literature [3, 12, 14].

DQM formats	Description
Ordinal	It indicates whether the data quality level is above average or below average, or it categorizes the DQ level as excellent, good, average and so forth.
Interval	It represents the DQ level using a 0-100 interval scale, where a higher DQ level indicates higher accuracy.
Probability	It represents the DQ level using a 0-1 probability scale which indicates the chances that the data are correct.
Range	It gives the lower and the upper limit where a specific data set.
Graphical	It uses colors to indicate the DQ level of a specific data set.

Table 2: Different DQM formats explored in literature

However, there are different issues in DQ tagging. First, there are no established rules, to the best of our knowledge, at which level DQM should be maintained

in databases. It is possible to have DQM at the level of the individual data item, at an attribute/column level and at the level of a relational table [4, 9]. However, the merits and demerits of these levels of DQM representations are not fully discussed in the literature. The most common level of DQM representation used by previous researchers is at the data item level [4, 9, 33, 40].

Second, determining the DQ dimension(s) for which quality measures should be stored as DQM is context-dependent. The most commonly used DQ dimension in the literature is the accuracy DQ dimension [4, 9, 33, 40]. This may acknowledge the importance of the accuracy dimension for different tasks [25]. This paper also uses the measure of the accuracy dimension as DQM in order to facilitate comparison with prior studies.

The third important consideration is the format of DQM, in particular how DQM is created, maintained and represented to the end users. The format in which DQM is represented can affect the decision making process and should be designed to facilitate the process [33, 43]. There are different DQM representations used in previous DQM research. Chengalur-Smith et al. [4] considered two approaches: 0-100 interval and n-level ordinal representation.

The n-level ordinal representation categorizes the DQ level as excellent, good, average, and so forth. The n-level ordinal representation could also be mapped into a two-point scale with a value of “above average” and “below average.” Fisher et al. [9] used the interval DQM representation. Shankaranarayanan et al. [38] used a percentage DQM representation where the quality level of the data represented with an 80% accuracy or completeness level is better than the quality of data represented by a 70% level. Moges et al. [24] conducted a pilot study to evaluate DQM representations by using two different types of DQM formats. These are DQM with lower and upper value limits (range representation) and probability representation. The range DQM format shows the minimum and maximum possible values for specific data where, for example, a specific data item can be in a range between 50-70. The probability DQM format presents the likelihood that the value of a specific data item represents its real value. Their pilot survey indicated the understandability of the probability DQM format. However, the probability and the interval DQM representation can be considered as similar with a minor distinction. On the other hand, Even et al. [8, 47] used a graphical representation of process metadata as an information product map (IPMAP) which uses colors to describe the quality level of data. Although there is no standard for DQM representations, many of prior DQM

researchers agreed on the understandability of an interval, a percentage and a probability DQM representation [4, 9, 24, 37]. As can be inferred from the text above, these three representations are similar.

The use of DQM for decision making purposes and the impact of DQM on the outcomes of decisions depend on the three issues discussed above. Therefore, the use of DQM should be investigated in consideration of those three elements. Although providing DQ metadata (DQM) along with the actual data set is considered to regain or complete the intuitive knowledge that is lost, it is important to identify whether decision makers are not complacent to the DQM (see Table 3). In addition, it is important to identify the impact of DQM on the outcomes of decision making. Including DQM in databases together with the actual data would be beneficial only if DQM is used and improves the decision outcome. As a result, many information systems researchers have responded to this need.

Chengalur-Smith et al. [4] investigated the use of DQM by using two DQM formats (ordinal and interval) and two decision strategies (conjunctive and weighted additive) (see Table 4). Their results indicated that when an ordinal DQM format was implemented, complacency was accepted for the conjunctive decision strategy but rejected for the weighted additive strategy for both simple and complex tasks. However, when the interval DQM format was implemented, complacency was rejected for both decision strategies for the simple decision task. Yet, complacency couldn't be rejected for both decision strategies for the complex task scenario which was further explained by the interaction effect of task complexity and DQM formats. The interval DQM representation gives detailed information about DQM unlike the ordinal DQM representation. To summarize, their results indicated an information overload. Information overload happens when the information given for solving a specific task is too much to be used by decision makers [2].

Fisher et al. [9], on the other hand, investigated how the experience of the decision maker, the available decision time, the time pressure and task complexity influence the use of DQM in decision making using an interval scale DQM representation. Their results indicated that complacency was rejected for experienced decision makers but couldn't be rejected for novices. However, these researchers haven't taken different types of decision making strategies into account.

Price et al. [33], investigated the use of DQM for two kinds of decision making strategies, namely additive and elimination by attributes (EBA) (see Table 4). Their results indicated that complacency couldn't be re-

jected for both decision strategies. However, decision time has been significantly extended.

Summarizing, Chengalur-Smith et al. [4] reported DQM use for the weighted additive decision strategy when an ordinal DQM is implemented for both simple and complex task scenarios. Yet, they reported DQM use for both decision strategies (weighted additive and conjunctive) when an interval DQM is implemented for only the simple decision task. On the other hand, Shanks and Tansley [41] reported DQM use for an EBA decision strategy when an interval DQM is implemented. However, Price et al. [34] couldn't reject complacency for both EBA and WA decision strategies.

Although the above studies reported the use of DQM in some circumstances, they didn't investigate the impact of DQM on the effectiveness of decision outcomes. In other words, those studies didn't report whether the presence of DQM positively or negatively impacted the outcomes of decision making. Yet, in contrast to all the above studies, Shankaranarayanan et al. [38] investigated the impact of DQM on the accuracy of the decision outcomes. They reported that DQM has a positive impact on the decision outcome.

Although many researchers have investigated the complacency of decision makers about DQM associated with many variables such as decision strategy, decision makers' experience and task complexity, to the best of our knowledge, with the exception of the study by Shankaranarayanan et al. [38], research on the impact of DQM on the effectiveness of decision outcomes lags behind. To fill this gap, this study analyzes the impact of DQM on the effectiveness of decision outcomes along with different variables such as the level of education, experience, DQ awareness, different decision making strategies and task complexity.

Mostly, prior DQM studies used an attribute level of DQ tagging, an interval DQM representation, two types of task complexity (simple and complex) and two types of decision making strategies in order to investigate the use of DQM for decision making purposes [4, 9, 23, 33, 41]. Additionally, these researchers define DQM usage in terms of the change in the preferred decision choice or the inclusion of a specific attribute in the decision processes.

Although two levels of task complexity (simple and complex tasks) and an attribute DQ tagging are implemented in this paper's experiment to enhance comparison with the above studies, our paper contributes a different method which is used to identify whether decision makers incorporated DQM given along with the data. In other words, the complacency of decision makers towards the DQM is clearly known in the decision

Decision outcome assessment	Description	How to measure
Complacency	Measures whether or not decision makers used the new variable, in this case the DQM, in their decision making process. If the decision outcome is similar for decision processes with and without DQM, then the decision maker is complacent to DQM.	Complacency measures whether the preferred choice is different between groups with DQM or without DQM.
Consensus	Measures the extent to which decision makers agreed on their decision making outcomes with and without DQM.	Consensus compares proportion of the best choice in both groups (with DQM and without DQM) such that the best choice for both groups may be different.
Consistency	Refers to the rankings of all alternatives from the most preferred to the least preferred.	Consistency can be measured by correlating the average rankings of each alternative for the two groups of decision makers, with and without DQM.
Efficiency	The time used by decision makers to accomplish a specific decision task.	Can be measured by comparing the difference between the time used in groups with DQM and without DQM.
Confidence	Measures the extent to which the decision confidence of decision makers is affected with and without DQM. It can be defined as the confidence level that decision makers have for the correctness of decision choices they made.	Can be measured by comparing the confidence level recorded for a specific decision task by different test groups.

Table 3: Measurements of decision outcomes discussed in DQM literature [4, 9, 39]

making strategies which subjects used to perform the decision task instead of a vague association of DQM usage in the change of preferred decision choices like previous studies [4, 9, 23, 33, 41]. Additionally, the paper measures the impact of DQM on the effectiveness of decision outcomes.

2.3. Relevant variables for the use of DQM

2.3.1. Decision making strategy

Many researchers have investigated different decision strategies in the past decades [42]. Simon [42] stated decision making as the process where decision makers can consider all attributes of all alternatives before selecting one. On the other hand, Nutt [28] identified four different decision making processes which aid decision makers to investigate different alternatives in order to choose the optimal one. These are analytical, judgmental, subjective and bargaining. As their name indicates, the judgmental, subjective and bargaining decision making processes depend mainly on the subjective analysis of the decision makers and their characteristics. Yet, the analytical decision making strategy allows the objective assessment of alternatives to reach to an optimal decision outcome [45]. In a structured task where a decision maker collects all relevant information, designs a decision scenario with a set of alternatives and chooses the optimal one, the analytical decision strategy is preferred because of its objectivity.

Payne et al. [31] explored three types of decision making strategies under the analytical decision making processes which are Weighted additive (WA), Conjunctive (CON) and Elimination by attribute (EBA). Table 4 shows the description of these decision strategies.

2.3.2. Experience

It can be reasonable to assume that experience is an important variable in decision making because it aids the decision process by incorporating life-time knowledge. Experienced decision makers can easily identify errors. They are also able to identify important aspects of a decision problem which may lead to a better decision outcome than inexperienced decision makers [18, 29].

An important consideration is the cognitive capacity limit of decision makers which might be positively affected by life-time knowledge [17]. The interaction between working-memory (WM) and long-term memory (LM) creates a cognitive capacity limit. WM stores data for a short period of time while LM stores data associated with life-time experiences. WM stores data for a while to merge with the data in the LM so that the specific decision task is conceptually represented. If the conceptual representation of the task is not enough to solve the decision problem, then WM draws data from the LM and applies logical rules to explore other conceptual representations of the decision task. This process will continue until the optimal solution for the task

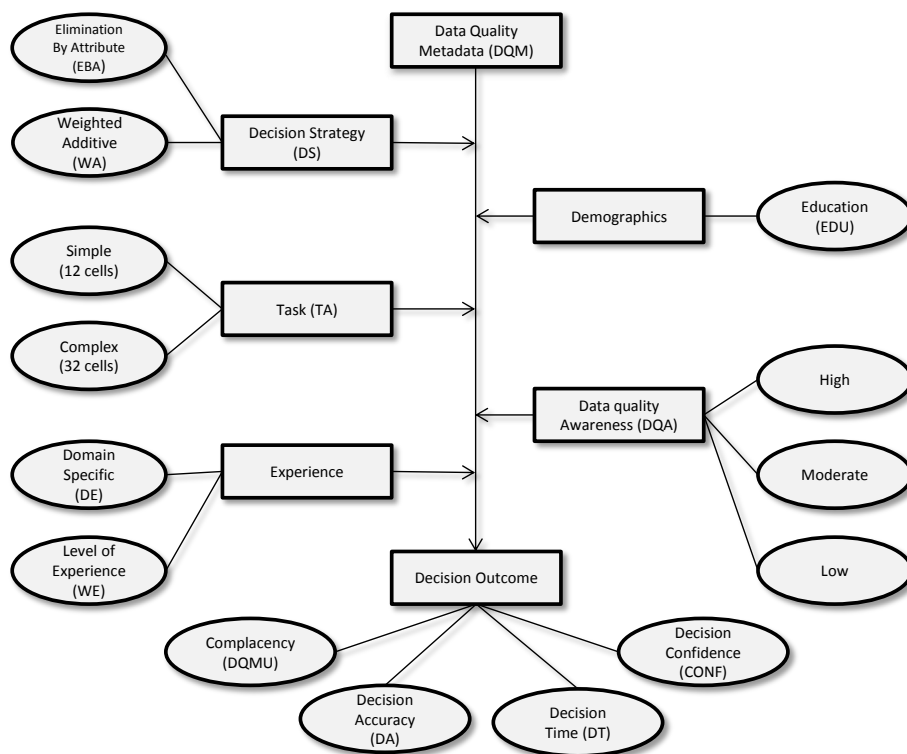


Figure 1: The research setup which shows the use and impact of DQM on decision outcomes, and its interaction with other variables.

Decision strategies	Description
Weighted Additive (WA)	It selects the highest sum value of the products of all criteria by their corresponding values. Let W_{1a1} & V_{1a1} and W_{2a1} & V_{2a1} , and W_{1a2} & V_{1a2} and W_{2a2} & V_{2a2} represent the weight and value of criterion 1 (C1) and criterion 2 (C2) for alternative 1 (A1) and alternative 2 (A2) respectively. The WA strategy would compare the value of $(W_{1a1} \times V_{1a1}) + (W_{2a1} \times V_{2a1})$ for A1 and $(W_{1a2} \times V_{1a2}) + (W_{2a2} \times V_{2a2})$ for A2 and selects the alternative with the highest value as an optimal solution of the decision task.
Elimination By Attribute (EBA)	It compares all alternatives based on the value of an attribute, the so-called deciding attribute, and then eliminates the alternatives which have lower values for the deciding attribute. The deciding attribute is usually selected based on its highest weight or decisive power for the decision task. If for example C1 is a decisive criterion because of its highest weight for a specific decision task, the EBA decision strategy would compare the values of $(W_{1a1} \times V_{1a1})$ and $(W_{1a2} \times V_{1a2})$ for A1 and A2 respectively, and selects an alternative with the highest value as an optimal decision outcome.
Conjunctive (CON)	It assigns a cut-off value for each criterion and selects an alternative with for each criterion a value above the cut-off value as an optimal decision outcome. The cut-off value is determined by business specialists where the value of it varies from task to task. For example, for some tasks a cut-off value of 50 for each criterion would be more appropriate but for other tasks it would be not sufficient. Mostly, an alternative with values above the cut-off for all or most of the criteria is selected as an optimal solution. Let the cut-off value for a specific decision task be 50 for each criterion. If V_{1a1} is 50 and V_{2a1} is 45 for A1 and V_{1a2} is 65 and V_{2a2} is 50, the CON strategy would select the alternative with the values for all or most of the criteria above 50. In this case, the optimal decision outcome would be A2.

Table 4: Summaries of the three kinds of decision strategies in literature [30]

is found or the capacity of the WM is exceeded [29]. Prior research reported that experienced decision makers can create a well-organized conceptual representation of decision tasks which can lead to an optimal solution compared to inexperienced decision makers [36].

However, it is not entirely correct to assume that experience is always good to arrive at the optimal solution for a decision task. Prior experience with the data may affect the feelings towards a specific data set. As a result, decision makers may rely more on their prior knowledge instead of using all the available data objectively. Therefore, decision makers may end the decision making processes early which, in turn, may negatively impact the outcomes of decisions [5]. Mao and Benbast [21] stated that experienced decision makers may consider their life-time knowledge more than the given information. Yet, they suggested that specialization may improve performance. For example, decision makers who had prior programming experience performed well in the given programming task.

Prior research indicated that when the level of experience increases, so will the use of DQM for decision making purposes. However, the level and type of experience are also found to have a different impact on the use of DQM. Fisher et al. [9] indicated that decision makers who have more managerial experience used DQM more but managers who have domain-specific experience used DQM to a lesser extent. In our work, we distinguish between domain-specific experience, where we assess whether the participant solved similar problems before and work experience, indicating whether the participant has any kind of general working experience.

2.3.3. Time

Decision time is a scarce resource for decision makers. Therefore, it needs to be utilized in a very efficient way. If providing DQM increases the decision time but doesn't increase the effectiveness of the decision output, then providing DQM can be assumed to have a negative impact on decision making.

Some researchers studied decision making with time pressure. Time pressure was measured differently by different researchers. Some researchers measured time pressure only by determining a specific time duration for a task [27]. However, other researchers differentiated between the time constraint and time pressure. They defined time constraint as a specific time allowed and time pressure was defined as a subjective reaction about the specific time allowed for performing a decision task [9]. Time pressure can happen whenever decision makers perceive the allowed time as not sufficient to complete the decision task [44].

Surprisingly, Fisher et al. [9] indicated that some decision makers may feel greater time pressure when they are given a longer time limit. In the same study, it is indicated that the time constraint didn't affect the use of DQM, while the time pressure was found to have a positive impact on the use of DQM. Decision makers who felt time pressure integrated DQM more than those who didn't feel time pressure in the same time constraint group. Conversely, a study by Price et al. [33] showed that providing DQM can significantly extend decision time.

In this study, the experiment didn't impose any time constraint on subjects. However, they were asked to register the time at which they started and finished working on the experiment.

2.3.4. Data Quality Awareness

One of the major goals of marketing is to make and maintain brand awareness. This is particularly important in an era when consumers actively search for information to assess their brand choices. Brand awareness is believed to have an impact on consumers' decision making, whereby the probability of brands being considered and selected can increase with the brand awareness level of the customers [19]. For example, customers are usually heard as being saying "I chose the brand I know," and, "I have heard of the brand so many times, I think it must be good."

Likewise, decision makers without DQ awareness may not fully use the DQM available in decision support systems. Similarly, DQ practitioners indicated the importance of creating DQ awareness to bring DQ problems into consideration [31].

Although at present many organizations are facing DQ problems, DQ awareness is often not considered thoroughly. One reason can be that the impact of poor DQ on organizations' performance is not clearly known. Although there is an intuitive feeling that DQ awareness improves the use of DQM by making decision makers alert, the impact of DQ awareness on the use of DQM and its effect on decision performance are not studied.

Therefore, in the experiment of this paper, the impact of DQ awareness on the use and effect of DQM on decision performance is investigated.

2.3.5. Task complexity

Task complexity can be determined by different factors, such as the amount of relevant information (the number of decision alternatives and attributes) and decision time available [31]. Task complexity increases with the amount of data that needs to be processed for a specific decision task [48]. Prior research has defined

task complexity using number of cells in the matrix of decision alternatives and decision criteria. A task which has 20 or less cells is categorized as a simple task, while a decision task with more than 20 cells is categorized as a complex task [31]. This study used this threshold to classify the task as simple or complex.

3. Research Methodology

The research methodology is developed by considering different factors which would have an impact on the use of DQM as depicted in Figure 1.

3.1. Research aim

Prior research indicated the importance of providing DQM along with the actual data so that decision makers can gauge the appropriateness of the DQ level for the task at hand [4]. However, there are two important questions which should be answered before deciding to include DQM into databases because of the associated cost of creating, maintaining and manipulating it. First, the question of whether DQM positively impacts the effectiveness (the accuracy/quality) of decision outcomes should be answered. Second, it should be clearly known whether decision makers are not complacent to the new information provided, in this case DQM. The latter question is adequately investigated in the DQ literature though there is no consensus about the results [4, 9, 34, 39, 40]. However, with the exception of a study by Shankaranarayanan et al. [38], the former has not been investigated sufficiently, to the best of our knowledge. Although both questions are important, the first one is more critical as the impact of DQM on the effectiveness of the decision outcomes can be either positive or negative.

In order to investigate the first research question, we employed three decision outcome measures, decision accuracy (effectiveness), decision confidence and decision time (efficiency). In other words, the impact of DQM on decision outcomes is measured in terms of those three dependent variables. Also, the interaction that DQM will have with other important variables such as experience level of decision makers and decision strategy is addressed under this research question.

H1 Decision makers who are less educated equally incorporate DQM into their decisions in comparison to decision makers who are more educated.

H1a Educated decision makers **include DQM more** into their decisions compared to less educated decision makers.

When the education level increases, the complacency to the new information decreases [9, 37].

H2 Decision makers who are less experienced equally incorporate DQM into their decisions in comparison to decision makers who are more experienced.

H2a Experienced decision makers **include DQM more** into their decisions compared to inexperienced decision makers. H2 and H2a are dealing with the general work experience level of the decision makers.

Experienced decision makers are less complacent to new information than novices [4, 9].

H3 Decision makers who have no domain-specific experience equally incorporate DQM into their decisions in comparison to decision makers who have domain specific experience.

H3a Decision makers without domain-specific experience (DE) **include DQM more** into their decisions than those with domain specific experience.

Specialization may prevent the use of all information objectively [9, 13].

H4 Decision makers who have prior DQ awareness equally include DQM into their decisions in comparison to decision makers who have no prior DQ awareness.

H4a Decision makers with DQ knowledge **include DQM more** into their decisions than those with no DQ knowledge.

Generally, decision makers are inclined to base their decisions on the known variables by ignoring the unknown variables [35]. Therefore, decision makers who have no prior DQ awareness are less likely to include DQM into their decisions for the reason that DQM is an unknown variable for them. In other words, the knowledge about DQ or DQ awareness triggers the use of DQM for decision making purposes. The latter lead to the following hypotheses:

H5 Decision makers who use a relatively simple decision making strategy to solve a decision task equally integrate DQM into their decisions in comparison to decision makers who use a relatively complex decision making strategy to solve a decision task.

More data and more choices may complicate the decision task. Decision makers can avoid cognitive overload by being complacent to additional variables such as DQM [32]. This can be explained by two well-known concepts in literature: information overload and cognitive capacity limit. Information overload can happen when a decision maker is asked to process more information than he/she is capable of. The second concept, the cognitive capacity limit, is the result of the interaction between working memory and long-term memory [2]. Processing more data and many alternatives may demand a high cognitive capacity level from the decision maker to hold and process the data used in the decision process. Therefore, solving a complex decision scenario may demand a high cognitive capacity unlike the simple decision task, resulting in some information being omitted. However, the feeling of information overload and cognitive capacity limit varies with the characteristics of decision makers. Prior research found that for the same amount of information novices may feel an information overload while experienced decision makers may not [6].

H5a Decision makers who use a relatively simple decision strategy (DS) **include DQM more** into their decisions than those who use a relatively complex decision strategy.

As reported above, if the decision process taxes cognitive capacity, decision makers tend to simplify the process by being complacent to new variables [17]. Prior research has indicated that the simplicity of a decision strategy can depend on the type of task [28]. A previous study by Moges et al. [26] identified the EBA decision strategy as relatively complex and WA as relatively simple because of its compensatory nature. Therefore, we expect that decision makers who use a WA decision strategy include DQM more into their decision process than those who use an EBA strategy.

The hypotheses from H1(H1a) to H5(H5a) are similarly assumed in both the simple and complex decision task scenarios. Therefore, in Section 4, Results and discussions, we use those hypotheses to present the results for both simple and complex decision environments.

H6 Decision makers facing either a simple or a complex task include DQM equally into their decisions.

H6a Decision makers facing a simple task include DQM into their decisions **more than** those who are assigned to solve a complex task.

3.2. Experimental setting

3.2.1. Pilot study

To determine the appropriate DQM representation, in a previous study [26], we have conducted a pilot experiment using two different types of DQM formats. These are DQM with lower and upper value limits (interval representation), and probability representation. The interval DQM format shows the minimum and maximum possible values for specific data. The probability DQM format presents the likelihood that the value of a specific data item represents its real value. Three groups were formed for the experiment, one that received interval DQM, one that received probability DQM and one without DQM. For experimental control, the decision strategies were limited to additive and EBA, and the experience level was limited to PhD students in Applied Economics.

The PhD students were randomly assigned to one of the three groups, each consisting of 10 students. A χ^2 statistical test indicated that there is no significant difference in the use of DQM between the interval and probability DQM format groups at the 95% confidence level. Thus, we conducted an exit-interview with the PhD students in the two groups (interval and probability) to find out (1) how they incorporated DQM in their decision-making, (2) how they understand the meaning of DQM and, (3) what kind of DQM format they would prefer. The interview analysis indicated that the PhD students in the interval DQM group didn't find an easy and uniform way of including the DQM in their decision-making process, which was also confirmed by a slightly higher time usage to finish the decision task compared to the other two groups. In contrast, the groups with the probability DQM format could easily understand the meaning of the DQM and used it all in a similar manner at the 95% confidence level.

Therefore, in the final experiment of this study, we employed a probability DQM format. In addition, we incorporated the usability study results that Price et al. [33] found. For example, we used the term accuracy for tag nomenclature and gave a detailed explanation of the meaning of the DQM with an example in the instruction section of the experiment.

Similarly, in this study, a pilot experiment with 10 PhD students was conducted in order to investigate the clarity and understandability of the experiment. In addition, the pilot study helped to deter-

mine the three decision strategies (Elimination by attribute (EBA), Weighted additive (WA), and Conjunctive (CON)) which subjects implemented to solve the decision tasks. The final experiment was designed to have the same decision solution using these three decision strategies so that the decision accuracy of each subject can be evaluated using a similar decision solution. Finally, the pilot study confirmed that there are no ambiguities in the experiment.

3.2.2. Final task

It has been indicated that the nature of the application domain used in DQM experiments may influence the extent to which DQM is used [33]. This is explained by the fact that participants may be less concerned that basing decisions on poor quality data negatively affects the decision outcome for a particular domain. Thus, previous research has suggested that the use of DQM for decision making should be investigated in different decision making environments, particularly in critical environments [33]. We therefore developed a new decision making environment which is a bankruptcy prediction task¹. This task is based on the Altman-Z model of bankruptcy prediction for non-manufacturing companies [1]. We employed all the four criteria ($\frac{\text{Retained earnings}}{\text{total assets}}$, $\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$, $\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$ and $\frac{\text{working capital}}{\text{Total assets}}$) to determine the financial health of a firm. The relative importance of each criterion is directly adopted from the Altman-Z model. The decision task was developed similarly to previous studies to facilitate comparison of the results [4, 10, 41]. Therefore, the decision making process in this study is described as the process of ranking firms according to their financial health from best to worst, based on the given criteria. In addition, we incorporated the usability study results that Price et al. [33] found in the experiment. According to the results of the pilot study in Section 3.2.1, the experiment was designed to have the same ranking result using any of the three decision mechanisms (EBA, WA and CON) (see Table 4) so that the ranking of the experiment can be evaluated using one correct answer. However, it was found that subjects in the final experiment used only the two decision strategies (EBA and WA) in their decision processes. Therefore, Figure 1 displayed only these two decision strategies.

The task was categorized into two types, simple and complex. The simple task asked subjects to rank the financial health of four banks based on the first

¹ see Appendix

three criteria ($\frac{\text{Retained earnings}}{\text{total assets}}$, $\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$, and $\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$), which has a total of 12 cells. Meanwhile, the complex task asked subjects to rank the financial health of eight banks based on the four criteria ($\frac{\text{Retained earnings}}{\text{total assets}}$, $\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$, $\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$ and $\frac{\text{working capital}}{\text{Total assets}}$), which has a total of 32 cells. Both the simple and complex tasks were further grouped into two types where some subjects receive DQM upfront and other subjects receive DQM later in the experiment. Subjects who didn't get DQM upfront with the experiment were provided with DQM later and asked if they would change their decisions because of the DQM provided. This is done particularly to increase the sample size for the complacency test. The four types of the experiment (simple task with upfront DQM, simple task with later DQM, complex task with upfront DQM, complex task with later DQM) were distributed randomly to subjects. Using a Friedman test at $\alpha = 5\%$, it was verified that no statistically significant differences exist in the DQM usage between subjects who were provided with DQM upfront and who were provided with DQM later on.

A clear description which explains the contents of the experiment, the meaning of each attribute and the expectations from subjects was also included in the experiment. The experiment was conducted in a controlled environment. An exit survey which consists of 28 questions was also conducted to gather demographic information after the experiment was finished. Finally, the subjects of the experiment were asked to register the time when they started and finished working on the experiment.

3.2.3. Participants

A total of 106 (80 business information systems students and 26 other students) subjects participated in the experiment. The participants have been further segmented based on the exit survey provided with the experiment. 60 students solved a decision task with DQM upfront and 46 students solved a decision task with DQM later. 42 and 64 participants solved the complex and simple decision task respectively. 30 of 106 participants have work experience. 35 have domain experience. 77 of the 106 participants have medium or high prior data quality awareness and 29 participants have no prior data quality awareness.

3.2.4. Variables

All the dependent and independent variables, and their descriptions and acronyms are included in Table 5.

3.3. Statistical Analysis

In order to test the significance of the obtained results, a number of statistical tests are applied in accordance with the literature. Each of the different tests is assessed at a significance level of 5% unless stated otherwise.

3.3.1. Chi-square

The χ^2 test basically investigates the null hypothesis, whether the frequency distribution of observed events in a sample is equal to an expected frequency distribution of the same events derived from a particular theoretical distribution or from the control groups in the observed samples. A χ^2 analysis can also be used to determine whether paired observations on two variables are independent from each other (e.g. the education level of subjects and their DQM usage). In this paper, a χ^2 test is used in this latter way to investigate the complacency level of the decision outcomes between different groups, to determine if there is a relationship between subject characteristics and DQM usage for decision making purposes [20].

3.3.2. Regression trees - Leave-one-out-cross validation

Regression trees

Tree-building algorithms generally define a set of logical environments by which different cases can be predicted or classified with some degree of accuracy. Regression trees are non-linear and non-parametric algorithms which predict continuous dependent variables using one or more continuous or categorical independent variables [15]. Regression trees are non-parametric and avoid the assumptions where tests such as Analysis Of Variance (ANOVA) and t-tests rely on, namely that data are normally, and independently and identically distributed (iid) [15]. In most cases, interpreting the results from the trees is very straightforward. Moreover, the trees do not assume any relationship (linear, non-linear or monotonic) between the predictor and the dependent variables. For example, decision accuracy can be negatively related to the use of DQM, but can also be positively related to the use of DQM if subjects have a high data quality awareness or high experience level whereby the tree can reveal such a non-monotonic relationship between the variables. Thus, linear regression trees are good methods for cases where there is little or no prior knowledge about the relationship between the dependent and independent variables. Therefore, they are particularly suited to analyze this experiment's data where there is no prior assumption about the three dependent variables i.e., decision accuracy (DA),

Variable	Description
Decision Accuracy (DA)	It is a discrete variable which is measured using the ranking of the firms for both complex and simple tasks. If a subject gets the ranking entirely correct, the accuracy is said to be 10/10. Otherwise, if a person gets one ranking wrong, the accuracy level decreases by 1.25 or 2.5 marks for the complex and simple task respectively. Although in a real life scenario, it is not always possible to a priori determine the optimal decision with uncertainties such as poor quality data, we designed the experiment to have a relatively correct ranking order by considering the DQ level given using the three commonly used decision strategies ^d in decision making processes. Therefore, the decision accuracy of each subject can be evaluated using a similar decision solution. As such, the definition of a 100% decision accuracy in this experiment's context is that one has achieved the predefined relatively correct ranking order without making any ranking error using one of the three decision strategies. In short, accuracy measures the number of errors that a decision maker makes in solving the decision task. The errors include calculation and inconsistencies in the decision making process.
Decision Confidence (CONF)	It is an ordinal variable which is measured with a 5-Likert scale which ranges from -2 to 2. -2 and -1 represent very low or low confidence about the decision outcome respectively, while 0 represents a neutral feeling about the decision task. Finally, 1 and 2 represent high and very high confidence about the decision outcomes.
Decision Time (DT)	It is a continuous variable which is measured by subtracting the start time from the finish time of the experiment. It is measured using minutes.
Data Quality Metadata Used (DQMU)	It is a categorical variable with a "yes" or "no" label. It is measured based on two different questions. First, subjects are asked to write the formula or method which they used to solve the task, if their method contains DQM as one decision variable and the decision solution belongs to the solution with DQM category, then the DQMU variable will get a "yes" label. Also, the response to the question "which variables were important to solve the decision task" is checked for its consistency to the solution provided. In almost all cases, the subjects' responses to the two different questions were consistent. Therefore, the use of DQM is correctly determined.
Data Quality Metadata (DQM)	It is a categorical variable with a "yes" or "no" label. It is measured based on the type of the experiment. For the experiment types which include DQM upfront, the DQM variable has a "yes" label, otherwise, it has a "no" label.
Decision Strategy (DS)	It is categorized into two broad types, Weighted Additive (WA) and Elimination by Attribute (EBA). Subjects are asked to explain which kind of decision strategy they used to solve the decision task. Based on their explanations, the decision strategy is classified into either EBA or WA.
Task Clear (TAclear)	It is a categorical variable which is measured by subjects' responses to the question "whether the experiment was fully clear". It has a "yes" or "no" label.
Task Type (TA)	It is a categorical variable with "Complex" or "Simple" label which is determined from the type of the experiment.
Data Quality Awareness (DQA)	It is a categorical variable which is measured by subjects' responses to the six basic data quality related questions asked in the exit survey. If a subject answers the four important questions ^{b c} correctly, a subject is said to have full data quality awareness. If a subject answers two or three of the four important questions correctly, then the subject is said to have medium data quality awareness. Subjects who answered only one question or did not answer any of the questions correctly, are said to have no data quality awareness.
Work Experience (WE)	It is a categorical variable which shows the work experience level of subjects. It is determined by the response to the question "how many years of work experience do you have" in the exit survey.
Domain Experience (DE)	It is a categorical variable which shows the level of previous similar experience with that of the experiment that subjects have. It is also determined by the response to the question "how many times did you solve this kind of exercise previously". If the answer to the question is zero, the person is said to be with no domain experience. Yet, if the answer to the question is 1 and above, the person is said to be with domain experience.
Education (EDU)	It is a categorical variable which shows the level of education that a subject has achieved.

Table 5: Summaries of the different variables in the experiment

^asee Table 4

^bHow do you define accuracy in the data quality context?

^cCan you please give one example of inaccurate data?

^dCan you please explain what is meant by data quality dimensions/attributes?

^eCan you please mention some data quality dimensions/attributes?

decision confidence (DC) and decision time (DT), and their predictors [16].

Leave-one-out cross-validation

Cross-validation is used to assess the performance of a regression or classification model on previously unseen data. Assessing the performance of the model is mainly the case in prediction analysis, where one predicts a model and determines its performance in practice. In general, model performance is measured by splitting the data in a training and test set. The model is estimated using the training set; the test set assesses its performance. However, in cross-validation, the original data set is split into several subsets, each of which is once used for testing purposes to assess the model's performance after the training phase. This technique helps to minimize overfitting and gives an insight on how the model will work on an independent real-life data set. In addition, it is very useful when a small sample size makes it difficult to split the data into separate training, validation and test sets. Leave-one-out cross-validation is one type of cross validation where each observation in the sample is used once as the validation data and the remaining observations as training data [11].

In this paper, we implemented a regression tree with leave-one-out cross-validation in order to predict the values of the three dependent variables (decision accuracy (DA), decision confidence (DC) and decision time (DT) from all the available independent variables.

3.3.3. Stepwise regression

A stepwise linear regression was also implemented in order to choose predictive variables with their interaction effect for each dependent variable, Decision Accuracy (DA), Confidence (CONF) and Time (DT). However, the results from the leave-one-out-cross validation and the stepwise regression are found to be similar with some minor distinctions. In addition, the results of leave-one-out cross-validation outperformed those of the stepwise regressions when the models were compared by the mean squared error (MSE). Therefore, in this paper, we will only present the results from the leave-one-out cross-validation tests [16].

4. Results and discussions

4.1. The use of DQM in decision making processes

4.1.1. Education, Experience, Data Quality Awareness and Decision Strategy - Simple decision task

As defined in Table 3, complacency measures the degree to which DQM is used in decision making pro-

cesses. A χ^2 test is conducted on the variable DQMU (see Table 5) in order to investigate the complacency level of different decision makers towards the DQM. The χ^2 test results are depicted in Table 6 and indicate that for the simple decision task, the complacency level about DQM of decision makers who are more or less educated is not significantly different at the 95% confidence level. Similarly, the complacency level between decision makers who have work experience and no work experience; and who have domain experience and no domain experience is not significantly different. In addition, there is no relationship between the complacency level and the decision strategies implemented. Therefore, H1-H3 and H5 are accepted, and H1a-H3a and H5a are rejected at the 95% confidence level for the simple decision task. Yet, there is a significant relationship between the complacency level of the decision makers and their DQA level at the 95% confidence level. In other words, subjects who have a high DQA integrated DQM more into their decision processes than subjects who have little or no DQA. Among 24 subjects who have little or no DQA, 17 of them didn't integrate DQM into their decision processes, yet, from 40 subjects who have a high DQA, only 15 of them didn't integrate DQM to solve the decision task. Therefore, H4 is rejected in favor of H4a at the 95% confidence level.

Simple Task					
Variables		DQMU	Obs.	Complacency	
EDU (H1)	Under graduates	Yes	20	$\chi^2 = 0.2591$	
		No	18		
	Post graduates	Yes	12		$p = 0.6107$
		No	14		
WE (H2)	No experience	Yes	23	$\chi^2 = 1.6967$	
		No	18		
	With experience	Yes	9		$p = 0.1927$
		No	14		
DE (H3)	Without DE	Yes	22	$\chi^2 = 0.2771$	
		No	20		
	With DE	Yes	10		$p = 0.5986$
		No	12		
DQA (H4)	Without DQA	Yes	7	$\chi^2 = 6.6667$	
		No	17		
	With DQA	Yes	25		$p = 0.00098^{**}$
		No	15		
DS (H5)	WA	Yes	18	$\chi^2 = 1.0667$	
		No	22		
	EBA	Yes	14		$p = 0.3017$
		No	10		

Table 6: The complacency level of different groups of subjects on their decision outcomes when Data Quality Metadata (DQM) is given and the decision task is relatively simple. $** = p < 0.05$.

4.1.2. Education, Experience, Data Quality Awareness and Decision strategy - Complex decision task

The χ^2 test results in Table 7 indicate that for the complex decision task, the complacency level about DQM of the decision makers who are more or less educated is not significantly different at the 95% confidence level. Similarly, the complacency level between decision makers who have work experience and no work experience is not significantly different. In addition, there is no relationship between the complacency level and the decision strategies implemented. Therefore, H1, H2 and H5 are accepted and H1a, H2a and H5a are rejected at the 95% confidence level for the complex decision task. Yet, the complacency level between subjects who have DE and who have no domain experience is significantly different at the 95% confidence level. Subjects who have prior DE on the decision task are more complacent towards the given DQM to solve the decision task than those subjects who have no domain experience. In other words, among 13 subjects who have prior DE, only 3 of them integrated DQM, yet, among 29 subjects who have no prior DE, 23 of them integrated DQM into their decision making processes. Similarly, there is a significant relationship between the complacency level of decision makers towards DQM and the DQA level of the decision makers at the 95% confidence level. In other words, subjects who have more DQA integrated DQM more than subjects who have no DQA, similar to the simple decision task. Among 5 subjects who have little or no DQA, 4 of them didn't integrate DQM to solve the decision task, yet, from 37 subjects who have a high DQA, only 12 of them didn't integrate DQM to solve the decision task, though both groups (subjects with DQA and subjects without DQA) were provided with DQM. Therefore, H3 and H4 are rejected and instead H3a and H4a are accepted at the 95% confidence level.

To summarize the results from Table 6 and 7, for both simple and complex decision tasks, the DQA level of decision makers have high impact on the degree to which decision makers are complacent towards the DQM. The higher the DQA level is, the more the decision makers integrate DQM into their decision making processes. This reinforces the suggestion from an earlier study by Fisher et al. [9] that organizations should conduct a seminar and DQM education prior to maintaining DQM in databases in order to fully benefit from it. Similarly, the results in Table 7 reveal that decision makers with domain experience used DQM less than those without domain experience for the complex de-

cision task. This confirms earlier findings by Gilliland et al. [13] and Fisher et al. [9] that more specialization may prevent the use of all the information objectively, whereby subjects with domain experience reportedly used DQM significantly less than those subjects without domain experience. On the other hand, the results in Table 6 and 7 indicate that there is no relationship between the level of education, work experience and the use of different decision strategies, and the complacency level of decision makers towards DQM. These results agree with the earlier findings that the education level and the type of decision strategy didn't affect the use of DQM for decision making processes [9, 34].

Complex Task				
Variables		DQMU	Obs.	Complacency
EDU (H1)	Under graduates	Yes	21	$\chi^2 = 0.1958$
		No	12	
	Post graduates	Yes	5	$p = 0.6581$
		No	4	
WE (H2)	No experience	Yes	22	$\chi^2 = 0.0808$
		No	13	
	With experience	Yes	4	$p = 0.7763$
		No	3	
DE (H3)	Without DE	Yes	23	$\chi^2 = 12.0361$
		No	6	
	With DE	Yes	3	$p = 0.0005^{**}$
		No	10	
DQA (H4)	Without DQA	Yes	1	$\chi^2 = 4.2262$
		No	4	
	With DQA	Yes	25	$p = 0.0398^{**}$
		No	12	
DS (H5)	WA	Yes	19	$\chi^2 = 1.2620$
		No	9	
	EBA	Yes	7	$p = 0.2613$
		No	7	

Table 7: The complacency level of different groups of subjects on their decision outcomes when Data Quality Metadata (DQM) is given and the decision task is relatively complex. $** = p < 0.05$.

4.1.3. Task type

As the results of the χ^2 test in Table 8 indicate there is no significant difference found between the complacency level of decision makers who solved the simple and complex decision task at the 95% confidence level. The well-known information overload theory would suggest that complex decision tasks would show more complacency to DQM than simple decision tasks. However, the results did not confirm this. Therefore, H6 is accepted and H6a is rejected. This result may be explained by the fact that both decision tasks (simple and complex) entailed a similar problem, with the task complexity determined only by the number of alternatives.

Complacency for the decision task when DQM is provided				
TA (H6)	Variables	DQMU	Obs.	Complacency
		Simple	Yes	
No	32			
Complex	Yes	26	$p = 0.2285$	
	No	16		

Table 8: The complacency level of subjects on their decision outcomes when Data Quality Metadata (DQM) is given in combination with the complexity of the decision task.

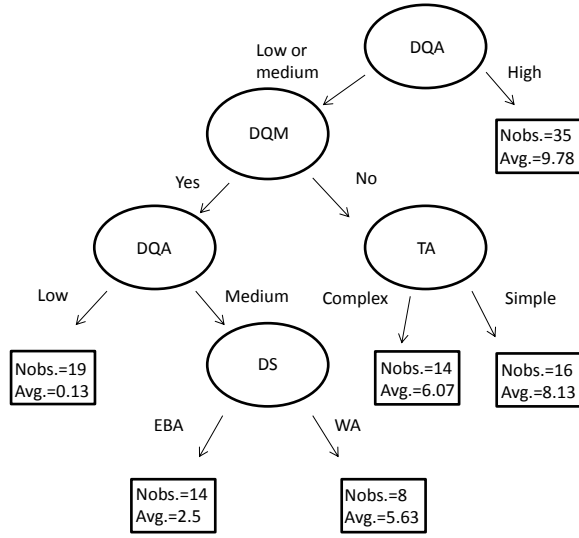


Figure 2: A regression tree for the decision accuracy (DA) with MSE=0.5715. The minimum score is zero and the maximum is 10.

4.2. Data quality metadata and its impact on decision outcomes

4.2.1. Decision accuracy

The regression tree in Figure 2 indicates that the dependent variable, decision accuracy (DA), can be predicted by the independent variables DQM, DQA, DS and TA with a low mean squared error of 0.5715. The DA variable is measured on a scale of 0 to 10, with the lowest value 0 and highest value 10. If decision makers have a high DQA, the probability of having a good level of decision accuracy is high. Yet, if decision makers have little or no DQA, decision accuracy depends on the presence of the DQM, DS and TA variables. In general, if decision makers who have no DQA integrate DQM into their decision processes, the decision accuracy will be very low. However, if decision makers who have moderate DQA integrate DQM into their decision processes and use a weighted additive decision strategy, their decision accuracy will be high. Conversely, if

decision makers do not integrate DQM into their decision processes, the decision accuracy depends on the the level of difficulty of the decision task, whereby a complex decision task leads to a lower decision accuracy. Summarizing the results, decision accuracy mainly depends on the level of DQA that decision makers have. Similarly, decision makers who have a high DQA use DQM more than those decision makers who have little or no DQA (see Table 6 and 7) and reach a high decision accuracy on their results.

In general, the results indicate that decision makers who have a high DQA can have high decision accuracy regardless of the decision strategy or the complexity of the task they dealt with. This may be explained by the fact that decision makers who have a high DQA are more educated as DQA knowledge is mostly acquired from the extra training in addition to the formal education.

4.2.2. Decision confidence

The regression tree in Figure 3 indicates that the dependent variable, decision confidence, can be predicted by the independent variables DQM, DQA, TAclear, EDU and DS with a MSE of 0.089. If decision makers do not integrate DQM into their decision processes and the decision task is not clear, the decision confidence is very low. Yet, if decision makers do not integrate DQM into their decision processes and the decision task is clear, then the decision confidence depends on the DQA level, as such, a high DQA leads to a better decision confidence. Similarly, if decision makers who have a high DQA integrate DQM into their decision processes, the decision confidence will be high. However, if decision makers who have little or no DQA integrate DQM into their decision processes, the decision confidence depends on the education level and decision strategy, whereby a high education level and an EBA decision strategy lead to a relatively high decision confidence. Summarizing the results, DQM usage, a high DQA level, the clarity of the decision task, a high educational level and a complex decision strategy have a positive impact on the confidence of decision makers on their decision outcomes. In other words, decision makers with a good prior data quality awareness and a high educational background, and who understand the decision task clearly will have a high decision confidence when they integrate DQM into their decision processes and use a more complex decision strategy such as EBA. Although Figure 3 depicts the interaction effects of different variables with DQM on the decision confidence and should be compared with other studies with interaction effect, the results can partially be compared with

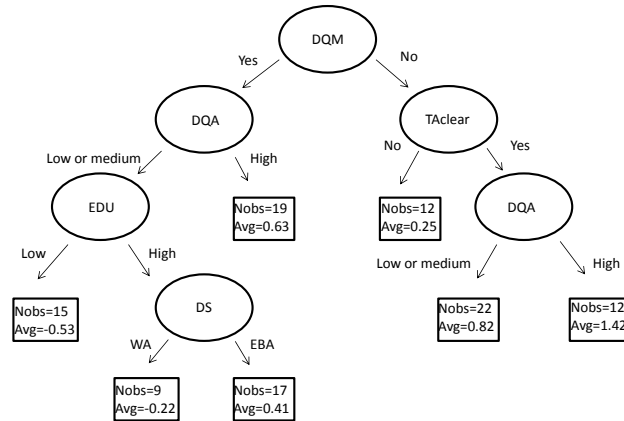


Figure 3: A regression tree for the confidence level of decision makers on their decision outcomes with $MSE=0.089$.

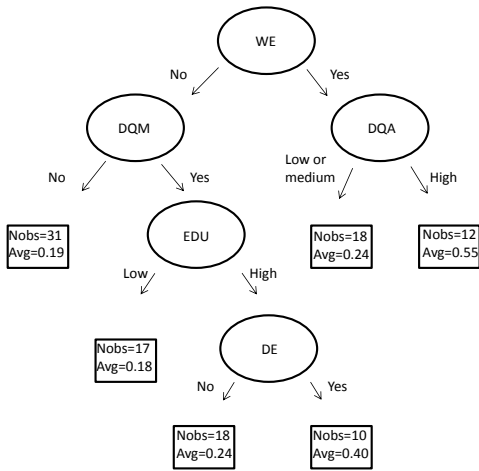


Figure 4: A regression tree for the decision time measured in minutes with $MSE=0.0118$.

the findings of Moges et al. [26] where the confidence level of decision makers was slightly higher for decision makers who used an EBA decision strategy to integrate DQM into their decision processes.

4.2.3. Decision time

The regression tree in Figure 4 indicates the prediction of the decision time in terms of the different independent variables. The results indicate that, in general, decision makers who have work experience will take more time to solve a similar task than those who do not have many years of work experience. Yet, for decision makers who have no work experience the decision time will depend on their integration of DQM into their de-

cision processes. Those without experience who do not use DQM will have a low decision time. Those without experience who do integrate DQM will vary their decision time based the education and the domain experience level of the decision makers. A high level of educational background and domain experience extends the decision time. In other words, a high DQA, a high educational background and experience level extend the decision time. One possible explanation would be that those decision makers who have a high DQA and educational background can integrate other knowledge or information from their prior encounters rather than limiting themselves only to the given information to solve the decision task. This can increase the decision time compared to novices who likely rely on only the given information to solve the decision task. In the same way, decision makers who integrate DQM can have extended decision time compared with those who didn't integrate DQM into their decision processes. This result confirmed previous findings by Price et al. [34] that the use of DQM can increase decision time.

5. Conclusion

In this study, a new experiment was conducted in order to investigate the use of DQM for decision making purposes and its impact on decision outcomes. The experiment was motivated by prior research results where there was no agreement on the use of DQM for decision making processes [9, 34, 40].

This study addressed different notions either suggested by previous studies or inferred from missing factors in their experimental designs. One of the suggestions addressed is that we created an equal or a similar

understanding of DQM among the participants by using a clear description and an example of what DQM represents which, consequently, helped to clear ambiguities. Also, the study incorporated all the variables studied in previous DQM research in addition to novel variables such as DQA which makes the study inclusive. This, in turn, helped to measure the effect of the variables on the use of DQM in a similar environment where similar subjects are used, consequently removing the impact of an experimental design.

Apart from the addition of the DQA variable, a second main contribution of this study is the way in which the decision outcome measures were defined and analyzed. First, the complacency of decision makers towards DQM, one of the most widely recognized decision outcome measures in the DQM literature, is inferred from three standards instead of association complacency indirectly with the change of the first decision choice or the usage of one attribute as in prior studies [4, 9, 34]. First, complacency is determined by participants' decision strategy where they explicitly indicated the formula they used to reach the decision solution. Second, it is derived from the category of the decision solution where the decision solution can either be categorized as a solution with DQM or without DQM. Finally, complacency is measured by the responses to the question "which variables were important in the decision processes". In nearly all cases, the three indicators of complacency were consistent for each subject. Second, the decision accuracy of each subject was evaluated using one possible decision solution. This could be possible because the experiment was designed to have one possible decision outcome for the three decision strategies (WA, EBA, and CON).

Another main contribution is that the way the three decision outcome measures, decision accuracy, decision time and decision confidence were analyzed. We used a tree-based algorithm to identify the impact of DQM and its interaction effect with other independent variables on these three measures. The results gave new insights on the impact of DQM on decision outcomes.

The use of DQM for decision making purposes was investigated using a χ^2 test. The results in Table 6, 7 and 8 indicated that the use of DQM is enhanced by prior data quality awareness of decision makers where decision makers with prior DQ awareness integrated DQM into their decision processes significantly more than those with no prior DQ awareness. One possible explanation would be that prior DQ awareness could increase the understanding of potential consequences of making decisions using incorrect data. Understanding the consequence of using flawed data encourages the

incorporation of DQM into decision processes. Another explanation can be, as prior market research indicated, that brand awareness increases the chance of the brand being purchased by users. Similarly, the more decision makers are familiar with the variable, the more they use it into their decision making processes [19]. Hereby, we can suggest that organizations should conduct a seminar or DQM training prior to maintaining DQM in datawarehouses. The results in Table 7 also indicate that the use of DQM decreases with domain experience level, whereby a high domain experience or specialization level is associated with a lower usage of DQM to solve the complex decision task. This result is consistent with the findings by Fisher et al. [9] that more specialization may avert the use of all available information. Additionally, decision makers who have more experience about the decision task may be more influenced by their prior experience than by other information given, such as DQM. On the contrary, education level, work experience level and decision strategy are found to have no effect on the use of DQM. Based on these results, we can suggest that special effort should be made to encourage users with domain experience to still make use of DQM.

The regression trees in Figure 2 suggest that in order to reach a high decision accuracy, DQM should be used by those who have prior DQ awareness. Yet, in general, the tree indicated that those who haven't used DQM reached a high decision accuracy though they solved a complex decision task. Put differently, although decision makers who have prior DQ awareness benefited from using DQM, decision makers who didn't use DQM could still reach a high decision accuracy. This can be explained by the preference of decision makers to use a complex decision strategy such as EBA to integrate DQM into their decision making processes. To clarify, among 38 subjects who used an EBA decision strategy, 26 integrated DQM into their decision processes.

The regression tree in Figure 3 indicates that, in general, decision confidence decreases when decision makers used DQM. Yet, decision confidence is found to be very high when decision makers have a high DQA. Similarly, when decision makers who have a high DQA solved the decision task using DQM, the confidence level is said to be high. In the same way, the regression tree in Figure 4 indicates that a high DQA and DQM use increased the decision time.

Although the decision accuracy and the decision confidence can be improved when decision makers who have a high DQA integrate DQM, the decision time increases. As DQ is contextual, for tasks which are critical and where the consequence of flawed data is high, pro-

viding DQM seems advantageous. Yet, for tasks which need a high efficiency, DQM seems rather counterproductive.

The general conclusion we can draw from this analysis is that DQM can be used and impacts the decision outcome when it is associated with certain characteristics of decision makers, some decision strategies and some task types. Therefore, it is clear that the cost of maintaining DQM can not be justified for all applications. Therefore, we suggest practitioners to investigate the benefits of DQM for their specific situation before justifying the cost of maintaining DQM in datawarehouses as DQ is indeed task-dependent.

Acknowledgements

This research was supported by the Odysseus program under grant G.0915.09.

References

- [1] E.I. Altman et al. Predicting financial distress of companies: Revisiting the z-score and zeta models. *Stern School of Business, New York University*, 2000.
- [2] H. Butcher. Information overload in management and business. In *Information Overload, IEE Colloquium on*, pages 1–1. IET, 1995.
- [3] P. Caplan. *Metadata fundamentals for all librarians*. American Library Association, 2003.
- [4] I.N. Chengalur-smith, D.P. Ballou, and H.L. Pazer. The impact of data quality information on decision making: An exploratory analysis. *IEEE Transactions of Knowledge and Data Engineering*, 11(6), 1999.
- [5] Janet M Dukerich and Mary Lippitt Nichols. Causal information search in managerial decision making. *Organizational Behavior and Human Decision Processes*, 50(1):106–122, 1991.
- [6] A. Edmunds and A. Morris. The problem of information overload in business organisations: a review of the literature. *International journal of information management*, 20(1):17–28, 2000.
- [7] L.P. English. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*, volume 1. Wiley New York, 1999.
- [8] A. Even, G. Shankaranarayanan, and S. Watts. Enhancing decision making with process metadata: Theoretical framework, research tool, and exploratory examination. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, volume 8, pages 209a–209a. IEEE, 2006.
- [9] C.W. Fisher and D.P. Ballou. The impact of experience and time on use of data quality information in decision making. *Information Systems Research*, 14(2):170–188, 2003.
- [10] C.W. Fisher, E.J.M. Lauria, and C.C. Matheus. An accuracy metric: Percentages, randomness, and probabilities. *Journal of Data and Information Quality (JDIQ)*, 1(3):16, 2009.
- [11] G.C. Cawley and N.L.C. Talbot. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition*, 36(11):2585–2592, 2003.
- [12] T. Gill, A. J. Gilliland-Swetland, M. Whalen, M. S. Woodley, and M. Baca. *Introduction to metadata*. Getty Research Institute, 2008.
- [13] S.W. Gilliland, L. Wood, and N. Schmitt. The effects of alternative labels on decision behavior: The case of corporate site selection decisions. *Organizational behavior and human decision processes*, 1994.
- [14] J. Greenberg. Understanding metadata and metadata schemes. *Cataloging & classification quarterly*, 40(3-4):17–36, 2005.
- [15] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. Springer, 2001.
- [16] SAS Institute. *SAS/STAT 9.22 User's Guide: Statistical Graphics Using ODS (Book Excerpt)*. SAS Institute, 2010.
- [17] D.E. Kieras and D.E. Meyer. An overview of the epic architecture for cognition and performance with application to human-computer interaction. *Human-computer interaction*, 12(4):391–438, 1997.
- [18] B.D. Klein, D.L. Goodhue, and G.B. Davis. Can humans detect errors in data? impact of base rates, incentives, and goals. *Management Information Systems Quarterly*, 21:169–194, 1997.
- [19] E.K. Macdonald and B.M. Sharp. Brand awareness effects on consumer decision making for a common, repeat purchase product: A replication. *Journal of Business Research*, 48(1):5–15, 2000.
- [20] N. Mantel. Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, 58(303):690–700, 1963.
- [21] J.Y. Mao and I. Benbasat. The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Management Information Systems*, 17(2):153–180, 2000.
- [22] D. Marco. *Building and managing the meta data repository*. Wiley New York, 2000.
- [23] H.T. Moges, K. Dejaeger, W. Lemahieu, and B. Baesens. Data quality for credit risk management: new insights and challenges. In *International Conference on Information Quality (ICIQ)*. University of South Australia, Adelaide (Australia), 2011.
- [24] H.T. Moges, K. Dejaeger, W. Lemahieu, and B. Baesens. A total data quality management for credit risk: new insights and challenges. *International Journal of Information Quality*, 3(1):1–27, 2012.
- [25] H.T. Moges, K. Dejaeger, W. Lemahieu, and B. Baesens. A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Journal of Information & Management*, 50(1):43–58, 2013.
- [26] H.T. Moges, W. Lemahieu, and B. Baesens. The use of data quality information (dqi) for decision-making: an exploratory study. In *Proceeding of the 2008 conference on Collaborative Decision Making: Perspectives and Challenges*, pages 233–244. IOS Press, 2008.
- [27] P. C. Morrow. Physical attractiveness and selection decision making. *Journal of Management*, 16(1):45–60, 1990.
- [28] P.C. Nutt. Types of organizational decision processes. *Administrative Science Quarterly*, pages 414–450, 1984.
- [29] J.W. Payne. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance*, 16(2):366–387, 1976.
- [30] J.W. Payne, J.R. Bettman, and E.J. Johnson. Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):534, 1988.
- [31] J.W. Payne, J.R. Bettman, and E.J. Johnson. *The adaptive decision maker*. Cambridge Univ Pr, 1993.
- [32] R. Price and G. Shanks. Representing data quality information

- usably. *Clayton School of Information Technology, Monash University, Technical report*, 243:1–19, 2009.
- [33] R. Price and G. Shanks. DQ tags and decision-making. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.
- [34] R. Price and G. Shanks. The impact of data quality tags on decision-making outcomes and process. *Journal of the Association for Information Systems*, 12(4):1, 2011.
- [35] S. Raghunathan. Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis. *Decision Support Systems*, 26:275–286, 1999.
- [36] A.H. Schoenfeld and D.J. Herrmann. Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5):484, 1982.
- [37] G. Shankaranarayanan and Y. Cai. Supporting data quality management in decision-making. *Decision Support Systems*, (42):302–317, 2006.
- [38] G. Shankaranarayanan and B. Zhu. Data quality metadata and decision making. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 1434–1443. IEEE, 2012.
- [39] G. Shankaranarayanan, M. Ziad, and R.Y. Wang. Managing data quality in dynamic decision environments: An information product approach. *Journal of Database Management*, 14(4):14–32, 2003.
- [40] G. Shanks. The impact of data quality tagging on decision outcomes. 2001.
- [41] G. Shanks and E. Tansley. Data quality tagging and decision outcomes: An experimental study. In *IFIP Conference on Decision Making and Decision Support in the Internet Age*, pages 399–410, 2002.
- [42] H.A. Simon. Administrative behaviour. *Australian Journal of Public Administration*, 9(1):241–245, 1950.
- [43] D.N. Stone and D.A. Schkade. Numeric and linguistic information representation in multiattribute choice. *Organizational Behavior and Human Decision Processes*, 49(1):42–59, 1991.
- [44] O. Svenson and A. Edland. Change of preferences under time pressure: Choices and judgements. *Scandinavian Journal of Psychology*, 28(4):322–330, 1987.
- [45] J.D. Thompson. *Organizations in action: Social science bases of administrative theory*. Transaction Pub, 1967.
- [46] R.Y. Wang and D.M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 1996.
- [47] Stephanie Watts, Ganesan Shankaranarayanan, and Adir Even. Data quality assessment in context: A cognitive perspective. *Decision Support Systems*, 48(1):202–211, 2009.
- [48] R.E. Wood. Task complexity: Definition of the construct. *Organizational behavior and human decision processes*, 37(1):60–82, 1986.

Appendix

In this section, the complex decision task is given as an example of the experiment.

Please write the hour and the minutes at which you started to complete this assignment in the following format (example 14:30 for 2:30 pm in the afternoon or 12:02 for 12:02 pm in the noon.....)

financial health or bankruptcy prediction Task Genet is working as a consultant in an accounting firm. She has been given 8 firms and was asked to rank them according to their solvency (financial health). Hence, she has begun the decision process of examining the firms. First, she identified four solvency determinant criteria and indicated the importance of each criterion using a weight based on an Altman Z-score model for non-manufacturing firms. The weight indicates the relative importance of each criterion in predicting the solvency of a company. The higher the weight, the higher the importance of a particular criterion in predicting the solvency of the company. Next, she represented the value for each criterion in euros, where higher values refer to more healthy firms. For example, a value of 90 euros for $\frac{\text{working capital}}{\text{Total assets}}$ ratio indicates a firm which is more healthy compared to a firm with a value of 50 euros.

Yet, she realized that the values may not be completely accurate as they are not consistent among different databases she checked. Thus, she decided to incorporate this uncertainty into her decision making process by using a [0, 1] accuracy measure where 0 indicates an inaccurate value and 1 indicates a perfectly accurate value. For example, an accuracy of 0.8 for a criterion's value indicates a 80% chance for the value to be correct.

However, because she is assigned to other work, she was unable to finish her ranking decision. Hence, her supervisor asked you to continue her work and to decide upon the ranking of the firms. You can assume that the accuracy, the value and the weight of the firms are correctly retrieved by Genet. Please rank the firms below according to their solvency from the most healthy firm (Rank 1) to the least healthy firm (Rank 8). Also, please explain why.

You can find the excel sheet of the experiment in the following link to facilitate your decision!
<http://www.econ.kuleuven.be/helen.moges/public/Experiment/ExperimentspreadsheetWDQM.xlsx>

Firm	Criterion	Accuracy	Value	Weight
Firm A	$\frac{\text{Retained earnings}}{\text{total assets}}$	0.8	84	3
	$\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$	0.8	24	1
	$\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$	0.5	80	7
	$\frac{\text{working capital}}{\text{Total assets}}$	0.5	16	6.5
Firm B	$\frac{\text{Retained earnings}}{\text{total assets}}$	0.8	20	3
	$\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$	0.8	16	1
	$\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$	1	48	7
	$\frac{\text{working capital}}{\text{Total assets}}$	1	30	6.5
Firm C	$\frac{\text{Retained earnings}}{\text{total assets}}$	0.4	100	3
	$\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$	0.2	80	1
	$\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$	0.1	90	7
	$\frac{\text{working capital}}{\text{Total assets}}$	0.1	80	6.5
Firm D	$\frac{\text{Retained earnings}}{\text{total assets}}$	0.6	52	3
	$\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$	0.8	48	1
	$\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$	0.8	54	7
	$\frac{\text{working capital}}{\text{Total assets}}$	0.8	26	6.5
Firm E	$\frac{\text{Retained earnings}}{\text{total assets}}$	0.7	76	3
	$\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$	0.8	24	1
	$\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$	1	55	7
	$\frac{\text{working capital}}{\text{Total assets}}$	1	40	6.5
Firm F	$\frac{\text{Retained earnings}}{\text{total assets}}$	0.8	24	3
	$\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$	0.5	18	1
	$\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$	0.2	20	7
	$\frac{\text{working capital}}{\text{Total assets}}$	0.2	56	6.5
Firm G	$\frac{\text{Retained earnings}}{\text{total assets}}$	1	50	3
	$\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$	0.8	40	1
	$\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$	1	50	7
	$\frac{\text{working capital}}{\text{Total assets}}$	1	28	6.5
Firm H	$\frac{\text{Retained earnings}}{\text{total assets}}$	0.7	52	3
	$\frac{\text{Market value equity}}{\text{Book value of total liabilities}}$	0.3	48	1
	$\frac{\text{Earnings before interest \& taxes}}{\text{Total assets}}$	0.3	51	7
	$\frac{\text{working capital}}{\text{Total assets}}$	0.2	29	6.5