

The Effect of Data Quality Tag Values and Usable Data Quality Tags on Decision-Making

Rosanne Price and Graeme Shanks

Department of Information Systems, University of Melbourne, Victoria, 3010, Australia
{rprice, gshanks}@unimelb.edu.au

Abstract. Prior research has shown that supplying decision-makers with *data quality (DQ) tags*, metadata about the quality of data used in decision-making, can impact decision outcomes in certain circumstances. However, there is conflicting evidence as to how or when decision outcomes are affected. In order to improve experimental soundness, the current research addresses possible sources of ambiguity in previous DQ tagging experiments with respect to DQ tag design. In particular, usability, semantics, and the rationale for assigning DQ values are explicitly considered in the DQ tag design of experiments. DQ tag values are designed to test the question of whether DQ tag impact is influenced by the relative importance of the “poor quality” attribute(s) to the specific decision task, thus potentially explaining discrepancies noted in observed results of prior research. The results showed no significant impact on decision choice, confidence, or efficiency; even with moderately rather than critically important attributes identified as being of poor quality. There was, however, some evidence of decreased consensus with DQ tags. These findings thus contraindicate the inclusion of data quality metadata in decision-making applications as a general business practice.

Keywords: Data quality tags, Decision-making, Contextual inquiry, Experimental soundness.

1 Introduction

Decision-making often relies on heterogeneous data sources that are remote from the user and possibly external to the organization or individual, particularly in the context of data warehouses. Decision-makers routinely use data whose context is unfamiliar and may vary depending on the individual data type and source. In particular, it is unlikely that the set of data used to make a decision is of uniform quality. The quality of the data used to make a multi-criteria decision potentially impacts the effectiveness of that decision. It has therefore been proposed [1] that information about data quality (DQ) be provided to decision-makers in the form of metadata, called *DQ tags*.

Decision outcomes could be influenced by the use of DQ tags. For example, decision makers may take time to consider data quality and reduce decision-making efficiency; some decision choices may be disregarded because of their low quality ratings; and the confidence in decisions may be eroded. For example, when using an

on-line system to search for a rental property; users may choose to ignore floor-space or treat it as a low priority if it is known to be unreliable compared to other criteria. Since the use of DQ tags is associated with significant overheads with respect to tag creation, storage, and maintenance; the adoption of DQ tagging as a business practice needs to be justified by a clear demonstration of its efficacy. The use of DQ tags by decision-makers may depend on factors such as decision-maker characteristics (eg. experience), the decision-making strategy employed, or the complexity of the decision-task (the number of decision criteria and alternatives). In order to predict the decision-making contexts most likely to benefit from the use of DQ tagging, it is therefore important to understand how such factors influence the use of DQ tags.

Decision-making strategies [2] can be classified based on whether multiple attributes of a single alternative or multiple alternatives for a single attribute are considered first (*alternative* or *attribute based* respectively) and whether desirable values for one attribute can or cannot compensate for undesirable values in another attribute (i.e. *compensatory* or *cutoff* respectively). An *Additive* decision-making strategy is therefore classified as *alternative* and *compensatory* because it involves assigning a desirability score to each attribute value for a single alternative, calculating a summed scores for each alternative, and then choosing the alternative with the highest sum—even though that alternative may have very undesirable values for some attributes. In contrast, the *Elimination by Attributes (EBA)* strategy involves elimination of alternatives not meeting the minimum requirements for the value of the attribute most important to the decision-maker. This process is repeated for other attributes in descending order of priority until only one alternative remains. Similarly, different levels of task complexity can be defined in terms of the number of decision criteria and alternatives and the ease of differentiating between alternatives.

1.1 Previous Work

Although DQ tags have been shown to impact decision outcomes given certain conditions [1,3,4], there is no agreement as to the necessary conditions (see [4, p.12]). Previous studies have in common the use of attribute-level tagging, the definition of two levels of task complexity (simple and complex), and the premise that a significant difference in decision-choice with and without DQ tags implies that the tags were used in the decision-making process when available. There is agreement [1,3,4] that increased task complexity is associated with reduced DQ tag usage, explained in terms of information overload. [4] reported DQ tag use only with an EBA but not an additive strategy, whereas [1] found DQ tag use to be more prevalent when DQ information was presented in a manner convenient for use with an additive strategy. In general, DQ tagging experiments to date indicate that DQ tag usage increases with experience (see [3, p.182])—with no use of tags by freshmen undergraduate students [3], use of tags for simple but not complex tasks for senior undergraduate students [1,4], and use of tags for both simple and complex tasks for professionals (ie. those with at least one year's prior work experience) [3]. All previous studies reported instances of decreased consensus in decision choice when DQ tags were used but no change in consensus when DQ tags were ignored (ie. present but not used).

Some studies have limitations with respect to the data sample size used and the control of the decision-making strategy employed by participants [1,3]. These studies relied on small paper-based data sets of fewer than 8 alternatives, in marked contrast to the large data sets characterizing on-line decision-making. This has implications for the *generalizability* (ie. applicability to the real-world and to contexts other than that of the experiment) of the experimental results to on-line decision-making.

In general, decision-making strategy was not controlled in [1,3] (except that cut-off scores were presented in some treatments in [1], see [5] for further discussion). Consequently, observed decision outcomes that were attributed solely to the use of tags could actually have depended partly or completely on the strategy or strategies used—potentially impacting experimental *validity* (ie. satisfactorily establishing that observed results are based on manipulations of dependent variables). In fact, evidence of the dependence of observed outcomes on decision-making strategy was reported in [4]. Concerns of scale and strategy were addressed in this study by using separate on-line interfaces, each with a different built-in decision-making strategy, to access an electronic database of 100 decision alternatives. However, usability concerns remain, as detailed in the following two paragraphs.

The specific DQ characteristic or criterion (e.g. *security* or *precision*) whose value is represented by the DQ tag is the tag semantics (i.e. meaning). The only guide to the meaning of the DQ tag used in previous experiments is its label (reliability in [1,3] and accuracy in [4]), without any further explanation given participants. In fact, a DQ tag could potentially be based on a number of different DQ criteria discussed in the literature [6,7,8,9,10]. Unless DQ tag semantics are specified explicitly, there might not be agreement between the interpretations of different experimental subjects or between their interpretations and that of the researcher. Since each individual finds his or her own internal interpretation clear, a problem with consistency is likely to go undetected in laboratory-based pilot studies. Such inconsistency could lead to random error in when or how DQ tags are used that impacts the *reliability* (ie. repeatability) of the experiment. Specific documented cases of such an occurrence in practice are discussed in [5], where data labelled as being of poor quality using the term *accuracy* were used for decision-making by participants who interpreted the DQ tag semantics as numerical precision and ignored by participants who interpreted the semantics as correctness (ie. the correspondence of the database to the real-world).

Tag usability is rarely discussed in the literature other than the use of pilot tests in [3,4], a technique whose limitations were illustrated above and are discussed in [11]. Since the use of DQ tags is not common in current business practice, there are generally no typical real-world precedents or widely understood conventions available to guide the researcher in designing or the user in understanding DQ tags. The only explicit consideration of different DQ tag representations is the comparison of DQ tag impact on decision-making using two-category ordinal (ie. *below average*, *above average*) versus continuous (ie. an integer between 1-100) representations of DQ information in [1]. Other possibilities for representing DQ values (eg. using ranges rather than single points, using graphics) and other aspects of DQ tag representation such as tag nomenclature or documentation have not been explicitly considered.

One previous DQ tagging experiment was designed to address these limitations and improve support for experimental *soundness* (i.e. *generalizability*, *reliability*, and *validity*) through explicit consideration of usability and DQ semantics [5]. The results

of that study differed from earlier DQ tagging research, insofar as there was no evidence of DQ tags impacting the preferred decision choice but some indication of decreased decision-making efficiency and consensus with DQ tags nonetheless. This was despite the fact that the decision task (ie. simple rental property selection), attributes, treatment group sample sizes, participant characteristics (ie. experienced), and DQ tag values were selected in order to be consistent with the treatments reported to have the highest levels of DQ tag use in earlier DQ tagging research [1,3,4]. In particular, the same attribute *commute time* had the lowest DQ value. However, participant comments regarding their concerns over the recent dramatic increase in petrol prices and environmental awareness highlighted the increased importance of this attribute in [5] as opposed to earlier DQ tagging studies [1,3,4].

Based on exit queries addressed to all participants with DQ tags, it was clear that the majority of participants with tags included *commute-time* in their decision-making because of its importance to the decision despite understanding that it was of low quality (ie. 28 out of the 33 participants with DQ tags). Thus this clearly showed that DQ tags are ignored for those attributes considered by decision-makers to be of critical importance. Consequent on this observation is two related questions as to whether (1) the difference in results between [5] and earlier studies [1,3,4] could be explained by a likely increase in the importance accorded the attribute tagged of lowest DQ (ie. *commute-time*), and (2) whether decision-makers would be more willing to ignore attributes designated to be of low quality if those attributes were perceived by decision-makers to be of lower importance. In order to answer these two questions, it was proposed that further experiments be conducted that have attributes perceived to be of less (eg. moderate rather than critical) importance tagged with the lowest DQ value. Thus, the goal of the current experiments is to test whether discrepancies in previously reported results can be accounted for by differences in the relative importance of the attribute(s) identified as being of poor quality.

The current experimental design has tag semantics, including derivation rules (the method used to calculate tag values), explicitly specified and is based on recommendations from a usability study conducted earlier. This helps to ensure that the DQ information provided is meaningful and the experimental context is credible. Section 2.1 overviews semantic and cost-based considerations in DQ tag design, section 2.2 overviews usability considerations in DQ tag design, and section 2.3 describes the basis for deciding which attributes to designate as being of low DQ in the experimental design. Section 3 presents the experimental design and research hypotheses. Results and discussion follow in section 4, with conclusions in section 5.

2 Design Considerations

2.1 DQ Semantics and Cost Considerations

The semantics and derivation of DQ tags for the experiment are to be explicitly specified in the experimental materials given to participants. The semantics of tags used in the current experiments are based on a previously defined information quality

framework called *InfoQual* [8], selected by virtue of its sound theoretical foundation and comprehensive coverage of different aspects of DQ. Different types of DQ tags can be defined based on *InfoQual*'s three DQ categories and their criteria. The first two categories (i.e. data *conformance* to rules and real-world *correspondence*) are relatively objective in nature since they are inherently based on the data set itself rather than the individual user or context of use. In contrast, the third category (i.e. *usefulness*) is necessarily subjective since it is based on context-specific information consumer views (see [8] for a detailed discussion). Therefore DQ tags based on subjective quality measures must be associated with additional contextual information (e.g. activity or task, organizational or geographic context, user profile) to be meaningfully interpreted or used. Within each of these three categories, *InfoQual* defines a set of individual DQ criteria further detailing DQ aspects. For example, the *usefulness* category includes criteria such as *security* and *timeliness*; whereas the *correspondence* category includes criteria such as *completeness*, *correctness*, and *consistency* (of the real-world representation in the database).

Since the creation, storage, and maintenance of tags incurs additional costs that offset potential benefits; it is desirable to restrict the scope to those choices that are likely to be the most practical in terms of simplicity, cost, and use. In the following discussion, we assume that a single decision-making *criterion* is represented as a database *attribute* or equivalently, assuming a relational database, as a *column* in a relational table; thus the three terms are used interchangeably. Cost-based issues that must be considered include those relating to DQ tag semantics, granularity, and level of consolidation, considered in detail in [11] and summarized here. Cost considerations thus suggest that DQ tags be:

- Objective: ie. have semantics based on objective rather than subjective aspects of DQ so that there is no need to store any contextual information. In the context of *InfoQual*, this means tags should be based on data *conformance* to database rules or real-world *correspondence* rather than on *usefulness*.
- Column-based: ie. specified in terms of a single DQ measurement for all the values of a single column. Relation, column, row, and field levels of granularity have progressively increasing information value but with the trade-off of increasing overheads. Column-level tagging is a natural compromise in the context of multi-criteria decision making, since the underlying cognitive processes involve evaluation of decision alternatives in terms of relevant criteria that are represented as columns.
- Consolidated: composite tags used when possible to logically combine related DQ criteria in order to limit storage overheads and reduce semantic complexity for users, eg. previous research [8, p.53] shows that users find it difficult to distinguish between individual criteria in the data *correspondence* category of *InfoQual*, preferring to combine them in a single summarized (i.e. category-level) concept instead—thus supporting the use of a single composite tag in this case.

These choices serve to limit the amount of extra information (and thus overhead costs) required for DQ tags. However, cost is not the only consideration for DQ tag design. As discussed in section 1.1, it is important to address usability concerns, especially given their potential impact on experimental soundness and the lack of common business conventions that could serve to guide experimental design or participant understanding of DQ tags. Usability questions relate to which aspects of

DQ and which possible DQ tag representations (including incorporation of such tags into the decision-making artefact) are the most understandable and relevant to decision-makers. Answers to the first question allow researchers to focus their attention and limited resources on those types of DQ tags judged by decision-makers to be the most likely to impact decision-making. Answers to the second question can be used to provide support for improved experimental soundness. The next section overviews a usability study conducted for this purpose.

2.2 Usability

Contextual inquiry [12,13]—the interrogatory component of the user-centered design approach called *contextual design* [12]—can be used to solicit user opinions on DQ tag design and use in the work environment rather than in a contrived laboratory setting. This technique involves on-site interviews of users—in this case decision-makers—while they demonstrate their real decision-making tasks. [12,13] describe how work processes and artefacts serve as reminders enabling decision-users to articulate their opinions in more detail and in reference to actual business practice (eg. rather than to the internal coherence of an experiment typical in a pilot test). Based on this approach, [11] describes in detail a usability study designed to solicit judgements from decision-makers as to the types and representations of DQ information most understandable and relevant in work and the current experimental context.

Following the contextual inquiry guidelines from [12,13] for a single work role (ie. a decision-maker), investigators conducted one-hour audio-taped interviews with 9 different decision-makers representing a diverse set of organizational (eg. sizes), data (eg. types and sources), decision (eg. domain), and technological (eg. software) contexts. Interviewees were asked to demonstrate a multi-criteria, data-intensive, and on-line decision-making task and to explain their experience in response to interviewer questions throughout the interview. In order to address specific usability questions related to DQ tag and experimental design, an additional and novel segment was added after the standard contextual inquiry interview. This ordering ensures that the additional segment does not bias the initial demonstration of current decision-making practice. In the additional segment, decision-makers were first asked to reflect on possible ways to improve the demonstrated decision-making task using DQ tags and then asked to review alternative DQ tag representations and a proposed decision-making interface incorporating tags. Transcribed interviews were analyzed as per specified in [12,13], with the end result being a table summarizing interviewee responses on a structured list of topics and a set of design recommendations based on this analysis. The value of these recommendations is supported by the general agreement between interviewed decision-makers despite their diverse contexts and despite the two different domains (work and experimental) considered.

The only type of DQ information considered to be of general interest at the attribute-based level was the degree of data *correspondence* to represented real-world values. Based on clear respondent preferences, the representation of such information would use the nomenclature *accuracy*, use a traffic light to graphically represent range-based tag values based on both color and position, and include explicit documentation of tag semantics and derivation. With respect to the proposed

experimental decision-making software artefact (ie. on-line decision-making interface), it was recommended that other interface elements (eg. attributes, value units) be documented using pop-up boxes and that desirability scores not be included in the decision-making interface. Such scores have been used in some earlier studies [1,3,4] to allow the relative desirability of different attribute values (ie. criteria) and alternatives to be compared despite differences in attribute measurement units (eg. dollars rent versus number of bedrooms for a rental apartment) and directionality (a lower rent but higher number of bedrooms is preferred); however, they were deemed to be unnecessary and confusing.

The usability study thus highlighted possible design issues in those earlier studies that did not explicitly consider usability [1,3,4]. In particular, their inclusion of desirability scores in the decision-making artefact and their use of a single numerical figure to represent DQ values without explicit specification of semantics is in direct contrast to the expressed preferences of the majority of interviewed decision-makers.

2.3 Selecting DQ Tag Values

The goal of the current experiment is to test whether decision-makers are more likely to ignore low quality attributes (ie. based on consideration of their DQ tags) if they are not of critical importance to the decision task (see section 1.1). Therefore, in contrast to all previous studies [1,3,4,5], the selection of new DQ values must take into consideration decision-makers' perceptions of the relative importance of different attributes to rental-property selection. In particular, we need to justify the decision as to which attribute(s) should be associated with the lowest DQ tag value. The challenge is to find an attribute of the right level of importance: neither too high (in which case, participants may ignore the DQ tags and use the attribute for decision-making regardless of its DQ tag value) or too low (since participants may ignore the attribute regardless of whether it has a tag or not because it is of no interest to the decision).

In order to find an attribute of *moderate* importance, we analyzed the rankings of decision attributes (ie. criteria) by priority from the experimental answer sheets completed by those participants not given tags in [5]. The decision attributes *rent*, *number of bedrooms*, and *commute time* were ruled out as possibilities based on their ranking, with more than 65% of participants ranking commute-time in the top two, more than 65% of participants ranking *number of bedrooms* in the top three, and all participants ranking *rent* in at least the top three and usually the topmost attribute(s) in terms of importance to the decision task of rental property selection.

Of the other two attributes, the majority (64% for *floor space* and 72% for *parking facilities*) ranked them as the two least important attributes. However, a minority of participants did consider these attributes to be important, as indicated by the fact that 36% ranked *floor space* as either the second or third most important attribute and 28% ranked *parking space* as one of the top three attributes in terms of importance. The contrasting figures (in contrast to *floor space*, *parking facilities* has a larger percentage of participants ranking it low but some rank it as most important) make it difficult to judge which is of more importance overall. However, it is clear that there is rarely a case (only 1 out of 39) where a given participant ranks both attributes as

high (ie. from the first to third most important attribute). Thus it is unlikely that a single participant will regard both attributes as so critically important that they will include both of them even if of poor quality. We expect that participants would be willing to remove one or the other if both are of poor quality. However, if participants rank both floor space and parking very low, then they may not include either floor or parking in their selections in any case—so making these attributes poor quality wouldn't affect apartment selections. Apartment selections can only be potentially affected by quality information if participants are interested in including at least one of these two attributes in their selected attributes without considering attribute quality. In fact, one of the decision-makers interviewed in the usability study expressed his opinion that DQ information was only of potential use for moderately important decision-criteria, since decision-makers would include critical criteria and exclude marginal criteria from their decision-making process in any case. The decision was therefore made to include both as attributes with the same and lowest DQ value, in the hope that at least one would be viewed as moderately important by participants who might then consider both the attribute and its DQ tag in their decision-making.

3 Research Design

In order to examine the impact of DQ tagging on decision outcomes for different decision-making strategies, we use a laboratory experiment. The research model is shown below in Figure 1, illustrating the causal relationships between theoretical constructs (represented in circles) using solid lines and the measurement relationships between theoretical constructs and their empirical indicators (represented using rectangles) using dotted lines. In common with [4], the focus is on multi-criteria, data-intensive and on-line decision-making with the decision strategy controlled. In general, therefore, we base our experimental methodology and design on theirs. However, central to our work—and distinguishing it from earlier DQ tagging research [1,3,4]—is the explicit emphasis on usability and DQ semantics in designing the DQ tags and incorporating them into a decision-making artefact. To this end, the experimental design in [4] is modified in accordance with the semantic-, cost-, and usability-based recommendations outlined in section 2.

The independent variables were decision-strategy and DQ tagging, both of which had two levels. Additive and EBA strategies are selected based on their contrasting properties (i.e. compensatory and alternative-based versus cut-off—therefore non-compensatory—and attribute-based respectively). DQ tags are either present or absent. The result is four separate experimental treatments: additive with DQ tags, additive without DQ tags, EBA with DQ tags, or EBA without DQ tags.

The dependent variables, potentially impacted by DQ tags, are decision complacency, consensus, efficiency, and confidence. As in previous DQ tagging studies [1,3,4], complacency and consensus relate to the impact of DQ tags on decision choice. Complacency refers to the degree to which decision-makers ignore DQ information. Thus a non-complacent outcome means that there is a significant change in the preferred decision choice when DQ tags are present as compared to that made when DQ tags are absent. The preferred decision choice is defined as that made

by the plurality of participants in the treatment group. Consensus refers to the level of agreement on decision choice between decision-makers: we are interested in whether the level with DQ tags is the same as that without DQ tags. Essentially, we are asking whether there is any difference in the number of participants making the preferred decision choice with and without tags, even though the preferred choice may not be the same for the two different groups. In accordance with the research model in [4], we investigate whether the presence of DQ tags significantly changes the time required to make the decision or the decision-maker's confidence in the decision choice made.

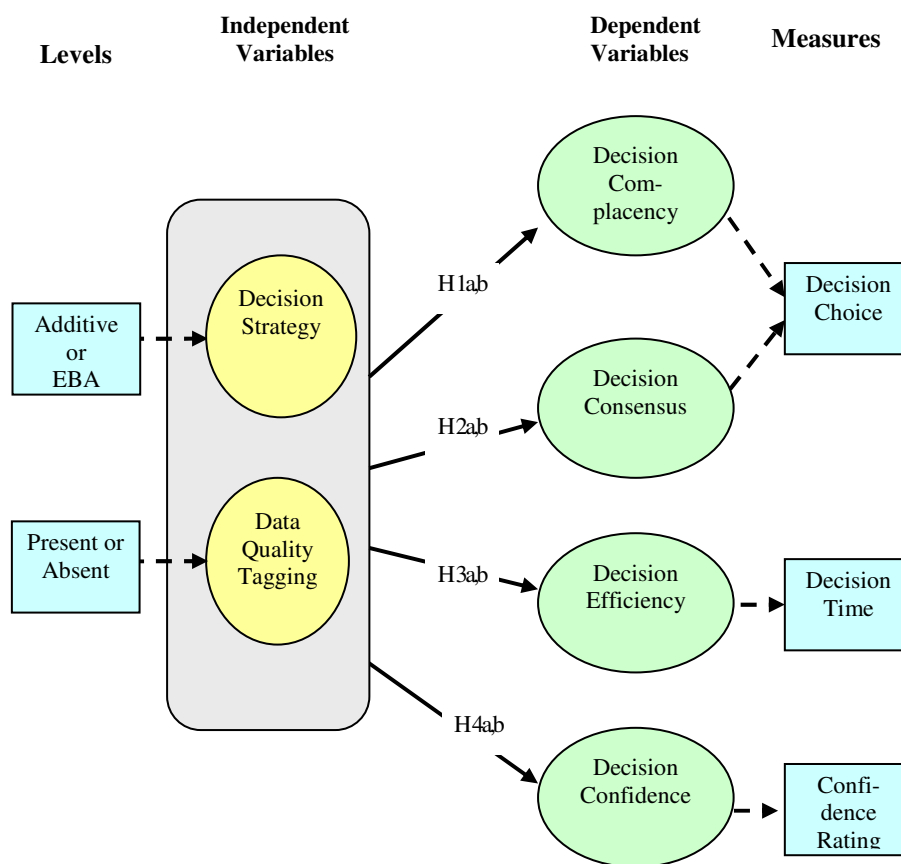


Fig. 1. Research Model

The case for the potential benefit of DQ tags would best be supported if the experiment shows that decision-makers are not complacent and have increased (or at least the same) level of consensus, efficiency, and confidence with tags. Such results require the rejection of the corresponding null hypotheses, formulated as follows: decision-makers are complacent (*H1*), and there is no difference in decision consensus (*H2*), efficiency (*H3*), or confidence (*H4*) with DQ tags for either the

additive (*a*) or EBA (*b*) decision-making strategies. A chi-squared statistic is used to test the first two hypotheses. Depending on whether the data is normally distributed or not, either an independent samples t-test or a Mann-Whitney test is used for the last two hypotheses. Tests were run using SPSS (version 15) and with reference to [14].

Since previous research shows that that DQ tag usage generally increases with decision-maker experience (see section 1.1) and considering the resources available, participants were limited to postgraduate university students enrolled in a masters or PhD degree. Such students are likely to have had more decision-making experience and prior professional experience than undergraduates. In fact, 52 of the 69 participants (ie. 75%) had prior work experience and 24 out of 69 (ie. 35%) had prior managerial experience.

The decision-making task used in the experiment involves selection of a preferred rental apartment based on the weekly rental cost, number of bedrooms, floor space, commute time, and parking facilities. The task involved recording decision start and finish times, nominating a confidence level using a 5-point likert scale ranging from *very low* to *very high*, and providing a brief explanation of the decision. In particular, participants were asked to write down which attributes—if any—they ignored in their search and why. Surveys of the target participant population of postgraduate university students showed that they typically had a good understanding of the domain and were frequent users of actual on-line rental property selection applications. The rental-property application domain, the set of attributes used (both their description and number), and the treatment group sample sizes are selected in order to be consistent with the simple task used in previous DQ tagging research [1,3,4,5], since prior research findings suggest that DQ tag usage is more likely in a simple rather than a complex decision-making task—especially for university students. Following the recommendations from section 2.3, the attributes *floor space* and *parking facilities* are both given the lowest (and equal) DQ values based on decision-maker perceptions that they were less important than other attributes in the rental property domain. In the context of the decision-making task, the above hypotheses are operationalized as no significant difference in the preferred apartment (*H1*), the number of decision-makers selecting the preferred apartment (*H2*), the decision time (*H3*), or the nominated decision confidence (*H4*).

We adopt a relational database-type interface and use Microsoft Access software for development as in [4] – both well-understood and widely used. Issues of scale and decision-making strategy (each potentially affecting the impact of DQ tagging) are similarly addressed through the use of separate on-line interfaces for each experimental treatment, each with 100 alternatives and a specific built-in decision-strategy and some including DQ tags. A set of instructions and an answer sheet were also developed for each different interface. In common with previous DQ tagging experiments, the alternatives in the databases are designed so that one apartment is clearly the most desirable without considering DQ information but is less desirable if DQ values from either or both of the low quality attributes are considered (ie. resulting in the associated low quality attribute(s) being ignored). In other words, the apartment automatically ranked as being the most desirable changes depending on whether both low quality attributes, only floor space, only parking facilities, or neither low quality attribute is included in the search. (This is true for either built-in decision-

making strategy; however, usually the top three ranked apartments change for the additive decision strategy.) The additive interface with tags is shown in Figure 2.

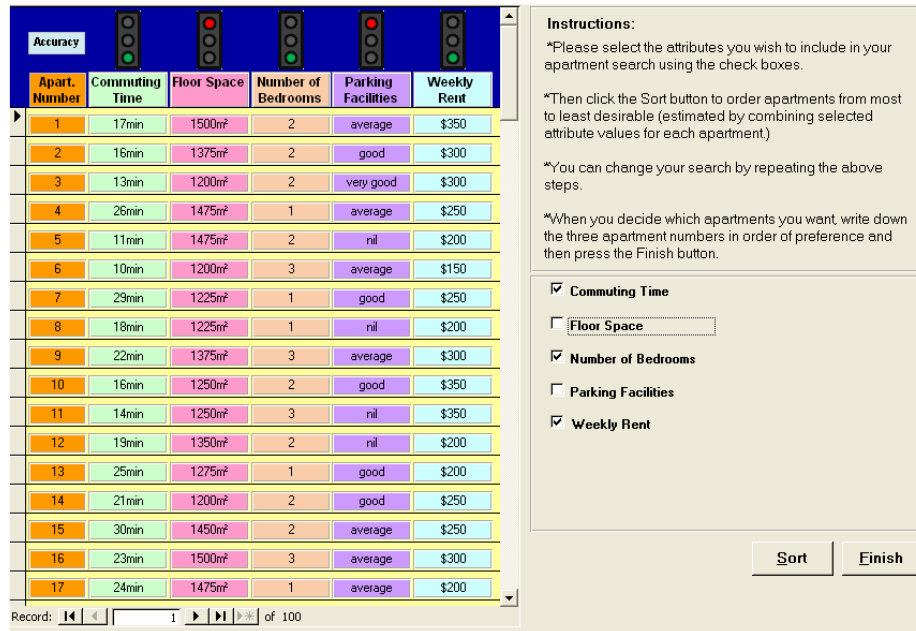


Fig. 2. Interface for Additive decision strategy with DQ tags

Calculated desirability scores are not displayed and concise explanations of interface elements are given in pop-up windows, in line with the recommendations resulting from the usability study described in section 2.2. Further to this purpose and in contrast with previous experiments, the current experiments use DQ tags with (1) semantics based on data correspondence (to the real-world), (2) range-based values represented by a traffic light symbol labelled *accuracy*, and (3) semantics and derivation explicitly specified¹ in both on-line and paper-based experimental materials. Furthermore, illustrative examples are included in explanatory notes given participants to clarify that the intended meaning of this term is *correctness* (based on real-world correspondence) rather than *numerical precision* (see Section 1.1 for documented evidence of such confusion).

The DQ tags used in the experiment are attribute-based, based on objective rather than subjective DQ information, and are consolidated. As explained in section 2.1, user preferences evident from prior research [8, p.53] support the consolidation of correspondence-based DQ information (described by the set of related *InfoQual* criteria in the *correspondence* category) into a single DQ tag.

¹ as “The traffic light symbol shows the percentage of column values in a random data sample that correctly corresponded to the represented real-world information, with a red light at the top indicating 0-33%, a yellow light in the middle indicating 34-66%, and a green light in the bottom indicating 67-100% of the values correct.”

5 postgraduate students and 1 professional participated in a pilot study of the experiment. They were asked to verbalize their thoughts during the experiment, resulting in minor modifications to the screen display and instruction wording and repair of one software bug. A further pilot study with 14 postgraduate students writing down their comments provided evidence that the materials were understandable and did not result in any new suggestions, indicative of saturation.

Participants were randomly assigned to one of 4 treatment groups. After being introduced to the experimental task scenario, decision-makers are asked to search for, select, and rank in order of priority their preferred rental apartment. Decision-makers select and—for the EBA strategy—rank those attributes they wish to include in the search. Desirability scores are automatically calculated and alternatives sorted in order of decreasing desirability based on the attributes selected and the specific decision-making strategy built into the interface. Participants can then repeatedly change their attribute selections and re-sort. For the treatments involving DQ tags, completed answer sheets were checked before participants left the laboratory to see if either of the attributes with the lowest DQ value—*floor space* or *parking facilities*—was used in the search. If so, the participant was asked whether he or she understood the meaning of the DQ tags and—if so—why they used it despite its poor quality.

4 Results and Discussion

Differences between an observed and the expected frequency distribution, where expected frequencies are derived from groups with no DQ tags, were checked with a chi-squared test. This test is non-parametric and therefore relatively free of underlying assumptions. Yates' Correction for Continuity is used as appropriate for a 2x2 chi-squared table. Results for analysis of decision complacency (*H1a*, *H1b*) and consensus (*H2a*, *H2b*) are summarized in Table 1. We can see that there is no significant difference ($p > .05$) in either for either decision-making strategy; therefore, it is not possible to reject the null hypotheses.

Table 1. Analysis of Complacency and Consensus

	Decision Strategy	
	Additive	Elimination by Attributes
Complacency	$\chi^2 = .000$	$\chi^2 = .896$
	$p = 1.000$ (<i>H1a</i>)	$p = .344$ (<i>H1b</i>)
Consensus	$\chi^2 = .000$	$\chi^2 = .896$
	$p = 1.000$ (<i>H2a</i>)	$p = .344$ (<i>H2b</i>)

For either decision-making strategy, Table 2 shows that the preferred apartment is the same with and without tags; so the chi-squared statistic is the same for complacency and consensus. For each treatment group, Table 2 also shows the total number and percentage of participants selecting any other than the preferred apartment under the label *Other*. Information about the plurality of participants next

in size compared to that of the participants selecting the preferred apartment is given under the label *Alternate*. The size of each treatment group is specified under the label *Total*. We can see that for all treatments, the plurality of participants selecting the preferred apartment is much larger (ranging from 11% to 64% larger) than any other plurality; thus the less sensitive non-parametric chi-squared statistic does not show a significant difference in consensus. However, for both decision-strategies, the percentage of participants selecting the preferred apartment decreases (from 41% to 35% with additive and from 76% to 56% with EBA) when DQ information is present (compared to when it is not). This suggests that there may be an overall decline in consensus with tags that is not detected by the chi-square test.

Table 2. Number of Participants Selecting Preferred and Other Apartments

		Number of Participants Selecting Apartment (% in treatment group selecting from set of specific apartments listed)	
		No Tags	Tags
Additive	Preferred	7 (41% for apt 83)	6 (35% for apt 83)
	Other	10 (59% for apt 29,57,70,90 or 91)	11 (65% for apt 36,70,90,91 or 98)
	Alternate	3 (18% for apt 57)	4 (24% for apt 70)
	Total	17	17
EBA	Preferred	13 (76% for apt 67)	10 (56% for apt 67)
	Other	4 (24% for apt 1,5 or 44)	8 (44% for apt 1,5,44 or 88)
	Alternate	2 (12% for apt 44)	5 (28% for apt 5)
	Total	17	18

An analysis of answer sheet and exit query responses showed that 23% (8 out of 35) of the participants given tags and 18% (6 out of 34) of the participants not given tags ignored *floor space*, either because they didn't care or—for 62% (5 out of 8) of those given tags who ignored the attribute—because the attribute was labelled as being of poor quality. For parking facilities, 51% (18 out of 35) of the participants given tags and 44% (15 of the 34) participants not given tags ignored the attribute, for the majority of participants because they didn't care but—for some (22%, 4 out of 18) of those participants given tags who ignored the attribute—because the attribute was labelled as being of poor quality. So we can see that only a minority of participants took the tags into account when determining which attributes to use in the search.

A visual inspection of relevant histograms, the Kolmogorov-Smirnov test, and the Shapiro-Wilks test revealed that both decision efficiency and confidence rating demonstrated significant variations from normal distribution for both decision-strategies. Hence, the non-parametric Mann-Whitney test was used for data analysis (*mean rank* and *level of significance* shown for each treatment group). The *mean* and *standard deviation* are also shown for each treatment group. Results for analysis of

decision efficiency (*H3a, H3b*) and confidence (*H4a, H4b*) are summarized in Table 3. There were no significant results, thus none of the null hypotheses can be rejected.

Table 3. Analysis of Time and Confidence

		Decision Strategy					
		Additive			Elimination by Attributes		
		Mean rank	Mean	SD	Mean rank	Mean	SD
Time	No Tags	16.75	8.59	5.48	18.53	5.53	2.50
	Tags	15.20	8.44	6.46	17.50	5.44	3.48
		p = .633 (H3a)			p = .763 (H3b)		
Confidence	No Tags	17.56	2.41	.71	18.32	2.00	.61
	Tags	17.44	2.35	.70	17.69	2.00	.84
		p = .970 (H4a)			p = .839 (H4b)		

5 Conclusion

Decision confidence remained *high* throughout all treatments, consistent with the findings reported in [4,5]. However, in contrast to these two studies, there was no change in decision time with DQ tags for either decision-making strategy.

With respect to decision complacency and consensus, the results of this study are markedly different from that of previous DQ tagging research [1,3,4] but consistent with recent experiments by the authors [5]. That is, there was some indications of an overall decline in decision consensus with tags even when decision-makers were complacent, i.e. did not change their preferred decision choice.

As detailed in section 1.1, current experiments were designed to answer questions raised in [5] by participant comments that even a very low quality attribute would be considered if it was very important for the decision task. The current experiments were therefore designed to see whether DQ tags would be used if a less important attribute was tagged as being of low quality; however, there was no change in results. Thus the majority of experimental participants from either decision-strategy treatment ignored DQ tags indicating that attributes were of poor quality, regardless of whether those attributes were regarded as important to the decision task or not. These findings do not support the adoption of DQ tagging in general business contexts.

Earlier DQ tagging studies [1,3,4] could be said to provide limited support for the possible utility of DQ tags in specific circumstances, although with caveats regarding the associated risk of increased decision-time and reduced consensus. However, we can see that their findings and recommendations were not reproduced in recent DQ tagging experiments even though the task complexity (i.e. simple) and target population (i.e. with decision-making, work, and managerial experience) selected for these experiments were based on the treatments reported in these early studies to be associated with the highest levels of DQ tag usage. A notable distinction in recent as

compared to earlier studies is their explicit emphasis on usability and semantics in experimental design, with the aim of providing better support for experimental soundness. We plan to conduct cognitive process tracing studies to explain these results more fully and to understand why DQ tags were largely ignored in these experiments. The experimental decision task and domain were selected to be consistent with previous research and familiar to participants; however, future studies are planned to test whether observed results generalize to other contexts.

Acknowledgments. We thank Ranjani Nagarajan and Ganesh A. for their excellent work in preparing materials and all their help and the Monash University Schools of Information Technology for letting us run some experimental sessions there.

References

1. Chengular-Smith, I.N., Pazer, H.L.: The Impact of Data Quality Information on Decision Making: An Exploratory Analysis. *IEEE TKDE* 11(6), 853–864 (1999)
2. Payne, J., Bettman, J., Johnson, E.: *The Adaptive Decision Maker*. Cambridge Univ. Press, Cambridge (1993)
3. Fisher, C., Chengular-Smith, I.N., Ballou, D.: The Impact of Experience and Time on the Use of Data Quality Information in Decision Making. *Information Systems Research*, 14(2), 170–188 (2003)
4. Shanks, G., Tansley, E.: Data Quality Tagging and Decision Outcomes: An Experimental Study. In: *IFIP Working Group 8.3 Conference on Decision Making and Decision Support in the Internet Age*, 399–410. Cork (2002)
5. Price, R., Shanks, G.: Data Quality Information and Decision-Making: Semantics, Usability, and Impact. Technical Report, Clayton School of Information Technology, Monash University (2009)
6. Eppler, M.J.: The Concept of Information Quality: An Interdisciplinary Evaluation of Recent Information Quality Frameworks. *Studies in Communication Sciences* 1, 167–182 (2001)
7. Even, A., Shankaranarayanan, G., Watts, S.: Enhancing Decision Making with Process Metadata: Theoretical Framework, Research Tool, and Exploratory Examination. In: *39th Hawaii International Conference on System Sciences (HICSS2006)*, 1–10. Hawaii (2006)
8. Price, R., Shanks, G.: A Semiotic Information Quality Framework: Development and Comparative Analysis. *Journal of Information Technology (JIT)* 20(2), 88–102 (2005)
9. Wand, Y., Wang, R.: Anchoring Data Quality Dimensions in Ontological Foundations. *CACM*, 39(11), 86–95 (1996)
10. Wang, R., Strong, D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems (JMIS)* 12(4), 5–34 (1996)
11. Price, R., Shanks, G.: Data Quality Tags and Decision-making: Improving the Design and Validity of Experimental Studies. In: *IFIP TC8/WG8.3 Working Group's International Conference on Collaborative Decision-Making (CDM'08)*, 233–244. Toulouse, (2008)
12. Beyer, H., Holzblatt, K.: *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann, San Francisco (1998)
13. Holzblatt, K., Jones, S.: Conducting and Analyzing a Contextual Interview (Excerpt). In: *Readings in Human-Computer Interaction: Towards the Year 2000*, 241–253. Morgan Kaufmann, San Francisco (2000)
14. Pallant, J. *SPSS Survival Manual A Step by Step Guide to Data Analysis using SPSS for Windows (Version 10)*., Allen & Unwin, Crows Nest, NSW, Australia (2001)