Abstract

There has been some evidence that decision outcomes can be impacted by providing metadata about the quality of the data used in a decision, but there is no agreement on the specific context in which such impact is observed. Furthermore, there has been little emphasis on the semantics and usability of such metadata in research to date, with consequent implications for experimental soundness. In this paper, we describe how these issues are addressed through the novel application, respectively, of the information quality framework InfoQual and the interaction design technique of contextual inquiry to the experimental design. In distinct contrast to previous studies, there was evidence that including data quality metadata can significantly increase decision time and may reduce consensus even though decision choice is not changed.

1. Introduction

One of the notable characteristics of decision-making in the electronic age is the decoupling of information production from use and the consequent loss of context. The problem is compounded by the need to integrate or synthesize data from different sources and contexts, as is typical of data warehouses or decision-support systems. In particular, the data is not likely to be of uniform quality. The quality of the data used can impact the effectiveness of a decision. It has thus been proposed [1] that metadata about data quality (DQ), called DQ tags, be provided to decision-makers. This could potentially impact decision outcomes if decision-makers have some discretion over the specific criteria to be considered or their relative priority. For example, the decision choice made may change if criteria that would otherwise be considered are disregarded because they are tagged as being of low quality.

Given the significant costs of DQ tags (eg. with respect to creation, storage, and maintenance), it is important to understand their impact on decision-making before considering their use as a general business practice. Previous investigators [1,2,3] have reported that DQ tags impact decision outcomes under certain circumstances, although there is conflicting evidence as to which specific circumstances [3, p12]. There is general agreement [1,2,3] that increased task complexity and reduced decision-maker experience is associated with reduced DQ tag usage, explained in terms of information overload. Decreases in consensus were reported only in conjunction with DQ tag use (ie. a change in decision choice when DQ tags were available). However, there is conflicting evidence as to how decision-strategy impacts DQ tag use. [3] reported DQ tag use only with an attribute-based strategy, whereas [1] found DQ tag use to be more prevalent when DQ information was presented in a manner convenient for use with an alternative-based strategy.

These 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 the preferred decision choice with and without DQ tags implies that the tags were used in the decision-making process when available. DQ tag usage is thus defined with respect to changes in preferred decision choice in the literature. We note, however, that there may potentially be some limited consideration of DQ tags by participants that affects decision consensus or efficiency even if decision choice is not significantly affected (ie. reported in the literature as “no evidence of tag use”). Although [3] addressed the limitations of [1,2] with respect to the size of the data sample used (only eight alternatives) and control of participants’ decision-making strategy (not constrained), none of these studies have focused on the semantics or usability of DQ tags.

Tag semantics (ie. meaning) relate to the specific DQ characteristic (eg. consistency) whose value is represented by the DQ tag. The only explanation of tag semantics given to participants in [1,2,3] was the label used (eg. reliability [1,2], accuracy [3]). A DQ tag could be defined using any of the DQ frameworks in the literature (eg. see [4] for early or [5,6,7,8] for...
recent frameworks). If the semantics of the DQ tags used are not explicitly defined and specified to participants, their interpretation may differ from that of other participants or the investigators. Such a problem may not be revealed by a pilot study. Individual subjects may say that the meaning of the tags is clear because they each have their own internal—even if erroneous—interpretation. This could lead to random error in when or how DQ tags are used that impacts experimental reliability (ie. repeatability) or validity (ie. showing that observations result from manipulation of dependent variables). Specific cases of such an occurrence in practice are documented in [9,10].

The only explicit reference to usability in [1,2,3] is the use of pilot tests in [2,3], a technique whose limitations were illustrated above and are further discussed in [9,10]. Given the novelty of DQ tagging in practice, we cannot depend on real-world precedents or widely-understood conventions to guide the design or use of such tags. Despite this limitation, the only explicit consideration of alternative DQ tag designs has been the comparison of how two-category ordinal versus integer representations impact DQ tag use in [1] (with inconclusive results). Other possible representations of DQ tag values (eg. using ranges rather than single points, using graphics) and other representation issues such as tag nomenclature or documentation have not been explicitly considered.

Previous papers by the authors [9,10] have argued for the explicit specification of DQ semantics and the explicit consideration of usability when designing DQ tagging experiments in order to improve support for experimental soundness (ie. generalizability, reliability, and validity). Tag semantics, including derivation rules (the method used to calculate tag values), must be specified explicitly and tag design must be usable in order to ensure that the DQ information provided is meaningful and the experimental context is credible. If participants do not understand the intended meaning of experimental materials, then this can lead to random error impacting experimental reliability. Lack of credibility can result in participant behavior that is inconsistent (ie. with real-world decision-making) or not a response to the experimental treatments (ie. if participants do not feel involved, they may be unaffected by the treatments), with consequent implications for generalizability (ie. applicability to the real-world and to contexts other than that considered in the experiment) or validity respectively (see [11, p265]). Such issues might have contributed to the previously noted inconsistency in the results of previous DQ tagging research with respect to when (ie. under what specific circumstances) DQ tags impact decision-making.

The importance of research intended to verify previous work (ie. with respect to the observations reported and the consequent interpretations propounded) and the relative paucity of such research in information systems (IS) literature is highlighted in [12]. The term extension is used to describe work that reproduces previous studies—but with key parameters changed—in order to improve understanding of the phenomenon under investigation and consolidate our knowledge by clarifying, confirming (or repudiating), or extending previous conclusions. In line with this view, the current paper describes DQ tagging experiments that focus on

1. areas of disagreement in [1,2,3], ie. the effect of decision-strategy on DQ tag use, and
2. experimental design issues relating to DQ tag semantics and usability not addressed in [1,2,3] that could potentially impact experimental soundness (and thus possibly be a factor contributing to such disagreement).

Usability issues are addressed based on the design recommendations of a study conducted by the authors for that purpose and reported in detail in [9,10]. To facilitate comparison of our experimental results with those of previous studies, we adopt the same definition of DQ tag usage in terms of changed preference and use the same decision task (ie. rental property selection for the simple task), poor-quality attribute (ie. commuting time for rental property selection), and level of DQ tagging granularity (ie. attribute-based DQ tags). We furthermore restrict our consideration to those specific experimental contexts consistently reported in previous experiments as having the highest level of DQ tag use, namely a simple rather than complex decision task and more experienced participants. This then allows us to define experimental treatments based on areas of disparity in previous research, as described above.

The rest of the paper is structured as follows. Section 2 discusses DQ tag semantics. Section 3 gives an overview of the usability study and consequent design recommendations in sufficient detail to understand the rationale for the current experimental design. Section 4 presents the experimental design and research hypotheses. Results follow in Section 5 and the conclusion in Section 6.

2. DQ Tag Semantics

DQ tag semantics could potentially be based on metadata either indirectly or directly related to DQ. Data characteristics such as source or processing
history do not directly describe DQ but are frequently employed by users as a basis for judging the likely quality (eg. trustworthiness) of data, as described in [5]. The semantics of tags directly related to DQ could be based on any of the DQ frameworks defined in the literature, eg. see [4,6,7]. As distinguished from other frameworks of comparable scope, the framework InfoQual [7,8] was selected because category definition and criteria classification both have a theoretical—thus rigorous—basis. (For example, criteria classification is an automatic consequence of category definition in InfoQual and thus consistent.)

Different types of DQ tags can be defined based on InfoQual’s three DQ categories and their criteria. The categories describe data conformance to defined rules, correspondence to the real-world, and usefulness for a given user and task. The first two categories are inherently based on the data set itself and thus relatively more objective than the third category (see [7] for a detailed discussion). Individual criteria in the usefulness category include timeliness and presentation. Requirements and preferences for such criteria depend on the specific data use and user. Therefore, DQ tags based on subjective quality measures must be associated with additional contextual information to be meaningfully interpreted or used. Since such contextual information would add considerably to the costs and complexity of keeping and using DQ tags, we restrict our consideration of possible DQ tag semantics to conformance and correspondence aspects.

The conformance category of InfoQual is unidimensional and consists of only the single criterion of conformance to data integrity rules. In contrast, the correspondence category has a set of individual criteria defined based on different cardinality constraints on mappings between the real-world and the IS (eg. complete means that each real-world instance must map to at least one IS element). Previous research [8,p.5] shows that users can find it difficult to distinguish between different mapping criteria, instead preferring to combine them in a single consolidated category-level concept. In this case, basing DQ tag semantics on individual mapping criteria would certainly increase storage overheads and potentially increase the semantic complexity of using DQ tags. Therefore, cost and complexity considerations suggest that we consider such DQ tag semantics based on the consolidated concept of data correspondence to the real-world rather than on individual mapping criteria.

In summary, the alternative types of DQ tag semantics considered in the usability study described in the next section include source, processing history, rule conformance, and real-world correspondence.

3. Usability Study

In order to improve the usability of experimental design for DQ tagging and thus better support experimental soundness, a usability study was conducted (see [9,10]). The specific goals of the usability study were (1) to observe relevant types of decision-making in practice, (2) to find DQ tag semantics and representations that decision-makers consider to be understandable and useful both in a general business context and in our proposed experimental context (ie. likely to improve their decision effectiveness or confidence), and (3) to obtain feedback on the understandability of the decision-making artefact proposed for use in the planned DQ tagging experiments.

Compared to other techniques for collecting usability judgements such as cognitive walkthroughs or participatory design workshops [13], the interaction design technique of contextual inquiry [14,15] was deemed the most relevant to the three goals listed above because the technique sources feedback from actual work environments, requires only a small sample of users, and can be applied outside the context of a single organization (see [15]). Contextual inquiry is the interrogatory component of contextual design, described in [14,15]. This technique involves interviewing decision-makers while they demonstrate their real decision-making tasks in their actual work environment. This approach is particularly suited to the goal of eliciting feedback on DQ tag design and use as relevant to current business practice (rather than to a single organization or to the artificial experimental context of a pilot test). Furthermore, the work context can serve as a reminder enabling decision-users to articulate their opinions in more detail [14,15]. We summarize here the design and resulting recommendations of the usability study reported in detail in [9,10].

Two investigators conducted one-hour audio-taped interviews with nine different decision-makers, as recommended by [14] for a single work role (ie. decision-maker). A diverse set of organizations (eg. size and type), data (eg. types and sources), decision domains, and technological contexts were represented. Decision-makers demonstrate and describe their reactions to a multi-criteria, data-intensive, and online decision-making task they use at work, prompted by interviewer questions throughout the interview. In order to address specific usability questions related to DQ tag and
experimental design without biasing initial reflections on work practice, an additional and novel segment was added after the standard contextual inquiry session.

In the additional segment, interviewees were asked how DQ information could be used to improve the demonstrated decision-making task. In particular, decision-makers were asked which type of DQ information (ie. DQ tag semantics) and which alternative representation (including DQ tag nomenclature, value, and explanation) were the most useful and understandable for their work. For example, alternative representations of DQ tag value for conformance and correspondence semantics considered included numeric point and range based representations and symbolic (eg. using traffic light symbols or minus and plus signs) or textual (eg. poor, OK, or good) range based representations. Participants were then shown the decision-making interface (based on [3]) considered for use in the current experiments. Participants were asked to reconsider their DQ tag design preferences (ie. semantics and representation) and to evaluate the understandability of the proposed decision-making artefact in the experimental context.

Transcribed interviews were analyzed by recording individual investigator interpretations on post-it notes and reconciling and structuring those interpretations as a team. The final result was 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 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 recommended representation of such information was as follows:

- use the term accuracy for tag nomenclature,
- use a traffic light to graphically represent range-based tag values using both color and position,
- include explicit documentation of tag semantics and derivation online, using brief explanations in pop-up boxes, and in any paper-based explanatory or instruction materials.

Furthermore, with respect to the experimental decision-making software artefact (ie. online decision-making interface) it was recommended that:

- other interface elements (eg. attributes, value units) be documented using pop-up boxes, and
desirability scores (included in the decision-making interface shown to participants and described in the next paragraph) be omitted as they were regarded as unnecessary and confusing.

Earlier DQ tagging experiments [1,2,3] have displayed desirability scores. Such scores allow the relative desirability of different attribute values and alternatives to be compared despite differences in measurement units (eg. dollars rent versus minutes commuting time for rental properties) and directionality (prefer a lower rent but more bedrooms). With respect to DQ tag design, most of these studies have used a single numerical figure to represent DQ information. None has used a graphical symbol or color to represent DQ information or has explicitly specified DQ tag semantics in the experimental materials used by participants. Thus prior experimental designs were not consistent with the user preferences expressed in the usability study.

4. Research Design

A laboratory experiment is used to examine the impact of DQ tagging on decision outcomes for different decision-making strategies. In common with [3], the focus is on multi-criteria, data-intensive and online decision-making. In general, our experimental methodology and design is based on theirs. Thus we use an online relational-type interface with a built-in decision-making strategy to access an electronic database with 100 alternatives. A separate interface is developed and used for each experimental treatment. However, in distinct contrast with earlier DQ tagging research [1,2,3], we focus on the usability and semantics of DQ tag design in order to provide better support for experimental soundness. Thus the usability recommendations outlined in Section 3 are used to revise the experimental design. As discussed in Section 1, additional considerations in designing the experiment were to:

1. define the experimental context to be consistent with those circumstances where DQ tag use was considered more likely based on results reported in [1,2,3], namely for a simple rather than complex decision task and with more experienced participants (ie. postgraduate and/or professional rather than undergraduate),
2. define experimental treatments to allow exploration of areas of disagreement in [1,2,3], namely how decision-strategy affects the impact of DQ tag use on decision-making, and
3. choose an experimental design consistent with [1,2,3] whenever possible (ie. given the above
constraints) in order to allow for a meaningful comparison of results. Thus, the application domain, the set of attributes used (both their description and number), the treatment group sample sizes, the DQ tag granularity, and the DQ tag values are selected in order to be consistent with the simple task used in earlier work [1,2,3].

Figure 1 shows the research model. The independent variables are DQ tagging and decision-making strategy. Each variable has two levels. DQ tags are either present or absent. The decision-strategy built into the interface is either additive or elimination by attributes (EBA). These strategies are selected as representative based on their contrasting properties [16], as explained in the next paragraph. The result is four separate experimental treatments: additive with DQ tags, additive without DQ tags, EBA with DQ tags, or EBA without DQ tags.

In the additive strategy, an overall desirability score is calculated for each alternative by summing the assigned desirability scores of its individual attribute values. Alternatives are ranked by comparing their overall scores: thus a high score in one attribute can compensate for a low score in another attribute for a given alternative. In contrast, the EBA strategy uses a hierarchical (ie. multi-level) sort. Alternatives are initially sorted based on individual desirability scores of the attribute most important to the decision-maker (ie. non-compensatory). Additional attributes are only considered (in order of importance) as needed to sort further those sub-groups of alternatives having the same value for the attribute previously used to sort. Thus the additive strategy is alternative-based (since all of the attributes for a single alternative are considered initially) and compensatory whereas the EBA strategy is attribute-based (since all of the alternatives are initially compared based on a single attribute) and non-compensatory.

The dependent variables are decision complacency, consensus, efficiency, and confidence. If the preferred decision choice remains the same with or without DQ tags, then the decision-makers are said to be complacent in that they ignored the DQ information. Conversely, a non-complacent outcome describes a significant change in the preferred decision choice with DQ tags. The preferred decision choice is defined as that made by the plurality of participants in the treatment group. To measure consensus, we compare the proportion of decision-makers selecting the preferred decision choice (which may be different for each treatment) with and without DQ tags. In the current context, efficiency refers to the time taken to make the decision. Confidence is the degree to which the decision-maker believes that he or she has made the best decision, measured in terms of a nominated confidence rating.

Based on these variables, 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 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 as compared to without DQ tags for either the additive (a) or EBA (b) decision-making strategies. The statistical analysis used is in accordance with standard statistical practice and recommendations (eg. see [17]) and previous DQ tagging studies. Thus a chi-squared statistic is used to test H1 and H2 since they involve a categorical dependent variable. Depending on whether the data is normally distributed or not, either an independent samples t-test or a Mann-Whitney test respectively is used for H3 and H4 since they involve a continuous dependent variable.

Given prior research evidence of increased DQ tag usage with more experienced decision-makers (see Section 1) and considering available resources, those university students most likely to have decision-making and professional experience (ie. postgraduate students enrolled in a masters or PhD degree rather than undergraduates) were used as participants. Of the 62 participants in this study, 25 (ie. 44%) had prior work experience and 10 (ie. 16%) had prior managerial experience.

The experimental decision-making task required participants to select preferred rental apartments based on the attributes shown in Figure 2. Surveys of postgraduate students showed that they were familiar with the task and were frequent users of actual online rental property selection applications. Participants were asked to note decision start and finish times, nominate a confidence level using a 5-point likert scale ranging from very low to very high, and briefly explain their decision. As part of this explanation, participants were asked if there were any attributes they ignored in their search and why.

In the context of the decision-making task, the hypotheses H1 through H4 are operationalized as no significant difference in the preferred apartment (H1), the proportion of decision-makers selecting the preferred apartment (H2), the decision time (H3), or the nominated level of decision confidence (H4) with as compared to without DQ tags.

An interface, a set of instructions, and an answer sheet were developed for each of the four different experimental treatments described previously. In
general, the experimental materials and DQ tag design used differ significantly from earlier experimental designs [1,2,3] in that they incorporate all of the usability recommendations from Section 3. However, in accordance with the third consideration described at the beginning of the section and in common with the rental property decision task from earlier studies [1,2,3], the attribute commuting time is tagged as having lowest DQ value (i.e. selected as the poor quality attribute) for those experimental interfaces with DQ tags. As in previous studies, the database alternatives are designed so that one apartment is clearly the most desirable without DQ information but is less desirable when DQ information is considered.

The additive interface with tags is shown in Figure 2, with a red light used to indicate that commuting time is poor quality and a yellow or green light used to indicate that other attributes are medium or good quality respectively. In this figure, the criteria to be considered have already been selected and the alternatives have been automatically sorted by decreasing desirability. In this case, desirability scores are calculated using an additive decision-making strategy and the selected criteria.

The experimental design was initially piloted individually, with 5 postgraduate students and 1 professional verbalizing their thoughts during the experiment. This resulted in minor changes to the
screen display and instruction wording and repair of one bug. A second pilot test with 14 postgraduate students found the materials clear and did not result in any new suggestions, indicative of saturation.

The assignment of participants to one of the 4 treatment groups was made randomly. Participants read instruction sheets and could ask questions before beginning the online decision task. The online decision task involved searching for and selecting their preferred rental apartment. Initially, participants select attributes to be considered in the search and—for the EBA strategy—rank the selected attributes in order of importance. 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. This process can be repeated with different attribute selections and/or rankings until the participant is satisfied and ready to make their apartment selection. For the treatments involving DQ tags, completed answer sheets were checked before participants left the laboratory to see if they used the poor quality attribute commuting time. If so, the participant was queried to see whether he or she understood the meaning of the DQ tags and—if so—why they used it despite its poor quality.

5. Results

A chi-squared test checks for differences between the observed and the expected frequency distribution, where expected frequencies are derived from groups with no DQ tags. This test is non-parametric and therefore relatively free of underlying assumptions [17]. Yates’ Correction for Continuity is used as appropriate for a 2x2 chi-squared table. Table 1 summarizes the results for the analysis of decision complacency (H1a, H1b) and consensus (H2a, H2b). Since no significant difference is shown for decision choice or consensus (p>.05) with as compared to without DQ tags, none of the corresponding null hypotheses (see Section 4) can be rejected.

Table 2 shows that the preferred apartment is the same with and without tags for either decision-making strategy; therefore, 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.

A comparison of preferred to alternate percentages within each treatment shows that the plurality of participants selecting the preferred apartment is much larger than any other plurality in every case; thus the less sensitive non-parametric chi-squared statistic does not show a significant difference in consensus. However, Table 2 shows a decreased percentage (30% less for additive and 15% less for EBA) of participants selecting the preferred apartment and an increased number (more than double the number) of different apartments preferred with tags as compared to without tags regardless of decision-strategy. This suggests there may be some decline in consensus with tags not detected by chi-square, although this apparent difference in consensus could be influenced by a perceived difference in the desirability of the apartment preferred with tags to that preferred without tags.

Based on participant responses to the exit query described in Section 4, only 4 (1 using the EBA and 3 using the Additive decision-strategy) of the 33 participants given DQ information ignored the low quality commuting time attribute. Of those, 2 didn’t care about commuting time and 2 ignored it because of its low quality. Of the 29 participants not given DQ information, 1 (for Additive strategy) ignored commuting time because they didn’t care about it. It is clear why there was no significant difference in decision choice with DQ tags when we consider that: (a) almost all (57 out of 62) of the participants used commuting time and that (b) very few (2 out of 33) of those participants who were informed that commuting time was of poor quality ignored it for that reason.

Of the 29 participants given DQ information but using commuting time in their search, only 1 did not understand the meaning of the tags (ie. share the intended interpretation of tag semantics). All the others said they considered commuting time to be so important that they included it despite its low quality. Petrol prices, environment, and traffic were all given as reasons for the attribute’s importance. The importance given this attribute is further highlighted by written comments from participants. Some participants rationalized their decision to consider commuting time despite its poor quality. Several participants commented that they would be visiting the selected apartments and so would be able to check the commuting time. Another said that the unreliability of commuting time was consistent with such property selection applications in practice, but “you had to use it [information about travel time] anyway because it was all that was available”. Others justified their behaviour by making assumptions about the degree of error (“even if commuting time is wrong, it is probably not too far off”).
Since decision efficiency and confidence showed significant variations from normal distribution for each decision-strategy (based on a visual inspection of relevant histograms, the Kolmogorov-Smirnov test, and the Shapiro-Wilks test), the non-parametric Mann-Whitney test was used for data analysis. Thus the mean rank and level of significance are shown with the mean and standard deviation for each treatment group. Table 3 summarizes the results for decision efficiency (H3a, H3b) and confidence (H4a, H4b). The only significant result (p=.046) is for decision efficiency using the additive strategy, where the presence of DQ information is associated with increased decision-time. Thus the null hypothesis is rejected, but based on decreased rather than increased decision efficiency.

Table 1. Analysis of Complacency and Consensus

<table>
<thead>
<tr>
<th>Decision Strategy</th>
<th>Additive</th>
<th>Elimination by Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complacency</td>
<td>$\chi^2 = 1.874$</td>
<td>$\chi^2 = .212$</td>
</tr>
<tr>
<td>p = .171 (H1a)</td>
<td>p = .645 (H1b)</td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>$\chi^2 = 1.874$</td>
<td>$\chi^2 = .212$</td>
</tr>
<tr>
<td>p = .171 (H2a)</td>
<td>p = .645 (H2b)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Number of Participants Selecting Preferred and Other Apartments

<table>
<thead>
<tr>
<th></th>
<th>No Tags</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>Preferred</td>
<td>12 (80% for apt 70)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>3 (20% for apt 77,83 or 98)</td>
</tr>
<tr>
<td></td>
<td>Alternate</td>
<td>1 each (7% each for apt 77, 83 and 98)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>15</td>
</tr>
<tr>
<td>EBA</td>
<td>Preferred</td>
<td>7 (50% for apt 5)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>7 (50% for apt 33 or 66)</td>
</tr>
<tr>
<td></td>
<td>Alternate</td>
<td>4 (29% for apt 33)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 3. Analysis of Time and Confidence

<table>
<thead>
<tr>
<th></th>
<th>Additive</th>
<th>Elimination by Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean rank</td>
<td>Mean</td>
</tr>
<tr>
<td>Time</td>
<td>No Tags</td>
<td>12.67</td>
</tr>
<tr>
<td></td>
<td>Tags</td>
<td>19.13</td>
</tr>
<tr>
<td></td>
<td>p = .046* (H3a)</td>
<td>p = .308 (H3b)</td>
</tr>
<tr>
<td>Confidence</td>
<td>No Tags</td>
<td>14.40</td>
</tr>
<tr>
<td></td>
<td>Tags</td>
<td>17.50</td>
</tr>
<tr>
<td></td>
<td>p = .305 (H4a)</td>
<td>p = .247 (H4b)</td>
</tr>
</tbody>
</table>
6. Conclusion

The current study focused on the impact of decision-strategies on DQ tag use, since this was a subject of disparity in results reported earlier [1,2,3]. Issues related to DQ tag usability and semantics were explicitly considered in the experimental design, with the aim of improving experimental soundness. The understandability of the resulting DQ tag design is demonstrated by the results of the exit query, which indicated that all but one of the respondents understood the intended meaning of the DQ tags.

Experimental design choices were made to ensure consistency with earlier research designs (eg. the same decision task, poor quality attribute, and DQ tag granularity were used to facilitate comparison of results) and to focus on those decision contexts consistently reported as being associated with the highest levels of DQ tag usage in earlier studies [1,2,3]. Despite this, the only result consistent with earlier work [3] was the high level of confidence reported regardless of treatment. With respect to decision complacency, consensus, and efficiency, the current study differed significantly from earlier DQ tagging research [1,2,3] in that (1) there was no evidence of DQ tag use with respect to changing the preferred decision choice and (2) there was a significant increase in decision time (with the additive strategy) and some indication (based on Table 2) of an overall decline in consensus without an associated change in decision choice. In contrast, previous studies [1,2,3] reported specific circumstances where DQ tags were associated with a change in preferred decision choice and that changes in consensus or decision time were only observed in conjunction with such a change.

The current study is distinguished from earlier studies [1,2,3] in that it does not offer even limited support for the possible utility of DQ tags. The only evidence of DQ tag impact on decision outcomes—reduced efficiency and consensus—is clearly detrimental to decision-making, thus contraindicating the general adoption of DQ tagging. Rather than resolving inconsistencies between reported results in [1,2,3] by confirming the results of one of the studies, the current study raises more questions. Possible explanations for the difference in experimental outcomes include the change in DQ tag design in the current study, the nature of the application domain considered, and temporal changes in the importance of the poor quality attribute to the decision.

Incorporating semantic and usability considerations in DQ tag and experimental design (and the subsequent impact on experimental soundness) may have contributed to the observed difference in results. The nature of the application domain selected in DQ tagging experiments to date may also influence the degree to which DQ tags are used. Since a poor choice of rental property selection may not be immediately obvious and does not ordinarily have serious consequences, experimental participants may be less concerned that using poor quality data negatively impacts the decision made. Finally, the importance of the attribute commuting time, marked as poor quality in all of the DQ tagging studies compared above, to selecting a rental property may have changed over time.

Even when available DQ information indicated that commuting time was of low quality, participant comments from the current study reveal that they considered the attribute much too important to ignore due to concerns about petrol prices, the environment, and traffic; Given that there has been a dramatic increase in petrol prices, environmental awareness, and traffic in the years since [1,2,3] were conducted, participants of the current study may have more reason to include commuting time than participants in earlier studies [1,2,3]. This suggests that it would be worthwhile to consider the effect of changing the attribute having low DQ values on DQ tag use, addressed in [18]. Whereas the current study keeps this parameter constant in order to facilitate comparisons with [1,2,3], [18] varies this parameter (using a different participant cohort) in order to explore the impact of DQ tag value on tag use. These two studies thus represent different extensions of earlier DQ tagging work [1,2,3], a research approach whose value to the investigative process is highlighted in [12].

In contrast to all other DQ tagging research reported to date [1,2,3,18], the current study shows that DQ tags can significantly increase decision time even when decision choice is not changed. It is clear from participant comments in the exit query that the majority of participants disregarded DQ information when deciding which criteria to consider and properties to select. This naturally raises the question as to why there was a significant increase in decision-time when DQ tags were apparently not used. One possible explanation would be that those participants given DQ tags spent some time considering whether and how to use the DQ information given, even when they eventually decided not to use the information. Thus the actual decision-making process could have been affected by DQ tags even though the decision choice did not change.

The current findings further suggest that consideration of DQ information requires more time for the additive than for the EBA decision-making strategy, since only the former showed a significant
increase in decision time with DQ tags. As proposed in [3] to explain observed time differences between strategies, this may be because it more difficult to assess how an individual attribute (eg. the poor quality attribute) affects the ranking of alternatives (eg. rental properties) with a compensatory (eg. additive) rather than a non-compensatory (eg. EBA) strategy. As acknowledged earlier in [2] but not yet addressed in DQ tagging research to date, further research is need to examine the impact of DQ tags on the decision-making process (rather than outcomes).

Cognitive process tracing studies are in progress to better understand how DQ tags affect the decision-making process and to help explain how DQ tags can impact decision efficiency without impacting decision choice. Plans are underway to test the generalizability of our findings in other decision-making contexts. In particular, future research should consider application domains such as emergency rooms where the potential consequences of basing decisions on flawed data are immediately obvious and more serious.

7. Acknowledgements

We thank the Australian Research Council for their funding and Ranjani and Ganesh for their help.

8. References


