Journal of the Association for Information Systems

JAIS

Research Article

The Impact of Data Quality Tags on Decision-Making Outcomes and Process

Rosanne Price
Monash University
rosanne.price@monash.edu

Graeme Shanks
The University of Melbourne
gshanks@unimelb.edu.au

Abstract

It has been proposed that metadata describing data quality (DQ), termed DQ tags, be made available in situations where decision makers are unfamiliar with the data context, for example, in data warehouses. However, there have been conflicting reports as to the impact of such DQ tags on decision-making outcomes. Early studies did not explicitly consider the usability and semantics of the DQ tag designs used experimentally or the impact of such tags on decision process, except in suggestions for future research. This study addresses these issues, focusing on the design of usable DQ tags whose semantics are explicitly specified and exploring the impact of such DQ tags on decision outcomes and process. We use the information quality framework InfoQual, the interaction design technique of contextual inquiry, and cognitive process tracing to address DQ tag semantics, usability, and impact on decision process, respectively. In distinct contrast to earlier laboratory experiments, there was no evidence that the preferred decision choice changed with DQ tags, but decision time was significantly increased and there were indications of reduced consensus. These results can be explained by understanding the impact of DQ tags on decision process using concurrent protocol analysis, which involves participants verbalizing thoughts while making a decision. The protocol analysis study shows that DQ tags are associated with increased cognitive processing in the earlier phases of decision making, which delays generation of decision alternatives.

Keywords: *Data Quality Tagging, Decision Making, Contextual Inquiry, Protocol Analysis.*

* Ping Zhang was the accepting senior editor. This article was submitted on 23rd December 2009 and went through three revisions.

Volume 12, Issue 4, pp. 323-346, April 2011

The Impact of Data Quality Tags on Decision-Making Outcomes and Process

1. Introduction

With the advent of data warehouses, there is a clear trend toward increasing dependence on data whose sources are varied or remote from the user and, thus, whose context is unfamiliar to the decision maker. Since the quality of the data potentially impacts the effectiveness of the decision, Chengular-Smith and Pazer (1999) have proposed that decision makers be given metadata with information about the quality of the data available, called data quality (DQ) tags. These tags could potentially be based on a number of different DQ categories or on criteria based on one of the DQ frameworks discussed in the literature (e.g., see Eppler (2001) for early or Nelson, Todd, and Wixom (2005) and Price and Shanks (2005) for recent frameworks).

As an alternative to providing DQ information directly in the form of tags, it has been suggested that decision makers could instead use process metadata (i.e., representing the history of how data was processed) to calculate DQ on-demand for a specific decision context based on an information manufacturing approach to DQ management (Shankaranarayanan, Ziad, & Wang, 2003; Shankaranarayanan & Cai, 2006). Shankaranarayanan, Even, and Watts (2006) further describe a non-experimental exploratory study of interactions between decision outcomes and user perceptions of process metadata usefulness, DQ, and decision-making efficiency. In contrast to a computational data processing-based approach to providing decision makers with DQ information, our focus in this paper is on DQ tags and tagging experiments.[1]

The process of creating, storing, and maintaining such tags is expensive and, thus, would need to be justified by a clear understanding of how DQ tags affect decision making. Several studies have investigated the impact of DQ tags on decision-making outcomes such as decision choice and consensus (Chengular-Smith & Pazer, 1999; Fisher, Chengular-Smith, & Ballou, 2003; Shanks & Tansley, 2002); however, they differ as to how and when decision making is affected. In particular, there is conflicting evidence about how decision strategy impacts DQ tag use. Shanks and Tansley (2002) report DQ tag use only with an attribute-based strategy, whereas Chengular-Smith and Pazer (1999) find DQ tag use to be more prevalent when DQ information is presented in a manner convenient for use with an alternative-based strategy. Shanks and Tansley (2002) further report an increase in decision time for an alternative-based decision strategy and a simple task even without evidence of DQ tag use. There is general agreement in these studies that increased task complexity and reduced decision-maker experience are associated with reduced DQ tag usage, explained in terms of information overload. Decreases in consensus are reported only in conjunction with DQ tag use (i.e., a change in decision choice when DQ tags are available).

These studies have in common the use of attribute-level tagging, consideration of two levels of task complexity (simple and complex) involving non-critical decision tasks, the focus on decision outcomes rather than process, and the definition of DQ tag usage in terms of changed decision choice. Thus, the assumption is that a significant difference in preferred decision choice with and without DQ tags implies that DQ tags are used in the decision-making process when available. Conversely, they presume that DQ tags are ignored when available if the resulting decision choice is not significantly different from that made without DQ tags. Shanks and Tansley (2002) address the limitations of Chengular-Smith and Pazer (1999) and Fisher et al. (2003) with respect to the size of the data sample used (only eight alternatives) and control of the decision-making strategy (decision-making strategy was not constrained). However, none of these studies has focused on the semantics or usability of DQ tags or the impact of DQ tags on decision process.

Tag semantics (i.e., meaning) relate to the specific DQ characteristic (e.g., consistency) whose value is represented by the DQ tag. Such semantics can be communicated explicitly through documentation or inferred based on the displayed DQ tag representation. The only explanation of tag semantics given to participants in the studies mentioned above was the label used (i.e., “reliability” (Chengular-Smith & Pazer, 1999; Fisher et al., 2003) or “accuracy” (Shanks & Tansley, 2002)). If

[1] Such an approach is consistent with decision makers’ preferences for a simple and easily understood representation of DQ information, as discussed in the Usability Study section of this paper.

the semantics of the DQ tags used are not explicitly defined and specified to participants, they may not agree on the interpretation of a DQ tag. Such a problem might 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 (i.e., repeatability) or validity (i.e., showing that observations result from manipulation of dependent variables). We document specific cases of such an occurrence in practice in a previous qualitative study (Price & Shanks, 2009).

Furthermore, if the DQ tag design (i.e., including both semantics and representation) is not understandable or relevant to experimental participants (i.e., usable), then participants may not find the experimental context or DQ tags credible. For example, decision makers interviewed in Price and Shanks (2009) have said that they would not trust DQ tag information unless the derivation method used to calculate DQ tag values were specified. Lack of credibility can result in participant behavior that is not consistent with real-world decision making or is not a response to the experimental treatments (Neuman, 2006, p.265), with consequent implications for generalizability (i.e., applicability beyond that of the specific experimental context considered) and validity respectively. Such issues might have contributed to the previously noted inconsistency in the results of previous DQ tagging research.

The only explicit reference to usability in the early DQ tagging studies mentioned above is the use of pilot tests in Fisher et al. (2003) and Shanks and Tansley (2002), a technique whose limitations were illustrated earlier. Despite the acknowledged importance of information presentation on decision behavior (e.g., in Chengular-Smith and Pazer, 1999) and the lack of widely understood conventions to guide the design or use of DQ tags, the only explicit consideration of alternative DQ tag designs is found in Chengular-Smith and Pazer (1999). They compare how two-category ordinal versus integer representation of DQ tag values impacts DQ tag use, with inconclusive results). Other possible representations of DQ tag values (e.g., using ranges rather than single points, using graphics) and other representation issues such as tag nomenclature or documentation have not been explicitly considered in previous DQ tagging experiments.

We published a previous paper (Price & Shanks, 2009) arguing for the need to conduct DQ tagging experiments that explicitly specify DQ semantics and consider the usability of DQ tag design in order to improve support for experimental soundness (i.e., generalizability, reliability, and validity). In terms of future work, Fisher et al. (2003) suggest that an investigation of decision-making process would be helpful to better understand the reported impact of DQ tags on decision outcomes. Berthon, Pitt, Ewing and Carr (2002) highlight the importance of research intended to verify or explain previous work (such as that suggested above) and the relative paucity of such research in information systems (IS) literature. As opposed to pure replication of previous research or pure generation of new research, 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 observations and conclusions. The change can be in the theoretical, methodological, or contextual basis of the experiment. In line with this view, the current paper describes first an extension of previous experiments investigating how DQ tags impact decision outcomes and then—in order to explain any observed impact—a research generation using qualitative methods to investigate how DQ tags impact decision process.

The research extension focuses on the effect of decision strategy on DQ tag use, since—as discussed previously—this is an area of disagreement in previous studies. Furthermore, this research extension explicitly considers DQ tag design issues of semantics and usability that were not previously addressed. As discussed earlier, these issues could potentially have implications for experimental soundness and, thus, may help account for differences in observed results.

In this paper, we first report on a laboratory experiment conducted to examine the impact of semantically specified, usable DQ tags on decision outcomes for two contrasting decision strategies.

We address semantic and usability issues based on the design recommendations of our previous study for that purpose (Price & Shanks ,2009). In order to understand any observed impacts of DQ tags on decision outcome, we then describe a cognitive process tracing study that investigates how the same DQ tags impact decision process by having participants verbalize their thoughts out loud as they make an online decision.

To facilitate comparison of the empirical work reported here with that of previous studies, we adopt the same definition of DQ tag usage in terms of changed decision choice preference and use the same decision task (i.e., rental property selection for the simple task), poor-quality attribute (i.e., commuting time for rental property selection), and level of DQ tagging granularity (i.e., 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 theoretical and methodological extensions to earlier experiments (Chengular-Smith & Pazer, 1999; Fisher et al., 2003; Shanks & Tansley, 2002) based on areas of disparity in results and on issues not previously addressed, as described above.

For those cases where there is a significant difference in decision outcomes observed with, as compared to without, DQ tags, we then use the qualitative technique of cognitive process tracing to address the as yet unexplored question of how DQ tags impact decision process.

The rest of the paper is structured as follows. The next section discusses DQ tag semantics. This is followed by an overview of the usability study and consequent design recommendations from Price and Shanks (2009) in sufficient detail to understand the rationale for the DQ tag design used in our research. The next two sections, respectively, present the design and results of the laboratory experiments examining the impact of DQ tags on decision outcomes. We explore the impact of DQ tags on decision process in the subsequent two sections, describing first the design and then the results of a cognitive process tracing study conducted for that purpose. The following section discusses our empirical results and relates the results of the experiment to earlier experiments and to those from the process tracing study. The final section concludes by considering the implications of these results for the use of DQ tags in practice and, based on the limitations of the current study, makes recommendations for future research directions.

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 (e.g., trustworthiness) of data, as described in Even et al. (2006). The semantics of tags directly related to DQ could be based on any of the DQ frameworks defined in the literature, e.g., see Eppler (2001), Nelson, Todd, and Wixom (2005), Price and Shanks (2005). As distinguished from other frameworks of comparable scope, we selected the framework InfoQual (Price & Shanks, 2005) because category definition and criteria classification both have a theoretical—thus, rigorous—basis. We briefly summarize here the InfoQual framework in sufficient detail to understand its use in the Usability study described in the next section.

With respect to tags whose semantics are directly related to DQ, 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 (e.g., employee bonuses must be less than 10 percent of their salaries), correspondence to the real world (e.g., the stored employee salary should match his or her actual salary), and usefulness for a given user and task (e.g., employee salary information is useful for the accounting department in order to issue paychecks). The first two categories are inherently based on the data set itself and, thus, are relatively more objective than the third category. Individual criteria in the usefulness category include criteria such as timeliness and presentation. Requirements and preferences for such criteria depend on the specific data use and user. Since a data set can be used for many different applications and users, each with different requirements, any measure of usefulness would be relevant only with reference to a particular context. Therefore, DQ tags based on subjective quality measures must be associated with additional contextual information (e.g., regarding user profile, task) to be meaningfully interpreted or used. Since such contextual information would add considerably to the costs and complexity of keeping and using DQ tags in practice, we restrict our consideration of possible DQ tag semantics to conformance and correspondence aspects.

The conformance category defined in 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 (e.g., complete means that each real world instance must map to at least one IS element). Price and Shanks (2005) note that end 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

Reported in detail in Price and Shanks (2009), a usability study was conducted to provide design recommendations for DQ tags experimentally and in practice based on decision makers' usability judgments. The goal was to observe relevant types of decision making in practice and to collect feedback on what DQ tag semantics and representation were considered to be the most understandable, relevant, and useful (e.g., likely to improve their decision effectiveness or confidence). We summarize the study here in sufficient detail to serve as context for the rest of the paper. We briefly describe the overall design of the study and those usability recommendations relevant to laboratory research (i.e., in order to provide better support for experimental soundness).

Compared to other techniques for collecting usability judgments such as cognitive walkthroughs or participatory design workshops (Benyon, Turner, & Turner, 2005), the interaction design technique of contextual inquiry (Beyer & Holtzblatt, 1998; Holtzblatt, Wendell, & Wood, 2005) was deemed the most suitable. This is 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 (Holtzblatt et al., 2005). 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-makersto articulate their opinions in more detail.

In line with contextual inquiry guidelines from Beyer and Holtzblatt (1998) and Holtzblatt et al. (2005) for a single work role (i.e., decision maker), investigators conducted one-hour audio-taped interviews with nine different decision makers representing a diverse set of organizational, data, decision, and technological contexts. Decision makers demonstrated a multi-criteria, data-intensive, and online decision-making task they use at work in response to interviewer questions. In order to address specific usability questions related to DQ tag and experimental design without biasing initial reflections on work practice, we added an additional and novel segment 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 (i.e., 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 (e.g., using traffic light symbols or minus and plus signs) or textual (e.g., poor, OK, or good) range-based representations. In order to obtain feedback on the understandability of an online decision-making artifact used experimentally for DQ tagging research, participants were then shown the decision-making interface from Shanks and Tansley (2002). Participants were asked to evaluate the understandability of the proposed decision-making artifact and to reconsider their design preferences in the experimental context.

We analyzed transcribed interviews per the guidelines in the sources mentioned above and produced 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 among 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 should use the nomenclature “accuracy,” use a traffic light to graphically represent range-based tag values using both color and position, and include explicit documentation of tag semantics and derivation. One participant commented that she “would not trust any DQ information supplied unless there was an explanation given…[which included the] derivation of such information.” Several commented explicitly that the representation used should be simple and immediately understandable. With respect to the proposed experimental decision-making software artifact (i.e., online decision-making interface), it was further recommended that desirability scores not be included in the decision-making interface. Such scores have been used in earlier DQ tagging experiments to allow the relative desirability of different attribute values (i.e., criteria) and alternatives to be compared despite differences in attribute measurement units (e.g., 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 earlier experiments that did not explicitly consider usability. In particular, the inclusion of desirability scores in the decision-making artifact and the use of a single numerical figure to represent DQ values without explicit specification of DQ tag semantics or derivation are in direct contrast to the expressed preferences of the majority of interviewed decision makers.

4. Laboratory Experiment: Research Method

We conduct a laboratory experiment to examine the impact of DQ tagging on decision outcomes for different decision-making strategies. In common with Shanks and Tansley (2002), 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. We develop and use a separate interface for each experimental treatment. However, in distinct contrast to earlier DQ tagging research (Chengular-Smith & Pazer, 1999; Fisher et al., 2003; Shanks & Tansley, 2002), we focus on the usability and semantics of DQ tag design in order to provide better support for experimental soundness. Thus, we use the usability recommendations outlined in the previous section to revise the experimental design. As discussed in the Introduction, 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 the three DQ tagging studies mentioned above; namely, for a simple, rather than complex, decision task

and with more experienced participants (i.e., postgraduate and/or professional rather than undergraduate).

2. Define experimental treatments to allow exploration of areas of disagreement in the DQ tagging studies mentioned above; namely, how decision strategy affects the impact of DQ tag use on decision-making.

3. Choose an experimental design consistent with the DQ tagging studies mentioned above whenever possible (i.e., given the above constraints) in order to allow for a meaningful comparison of results. Thus, we selected 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 to be consistent with the simple task used in earlier research.

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 value of the DQ tag associated with a given attribute describes the quality of that attribute. Consistent with previous DQ tagging experiments, one attribute is specified to be of much lower quality than all of the other attributes. We refer to this attribute as the poor quality attribute. The decision strategy built into the interface is either additive or elimination by attributes (EBA). We selected these strategies as representative based on their contrasting properties (Payne, Bettman, & Johnson, 1993), 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.

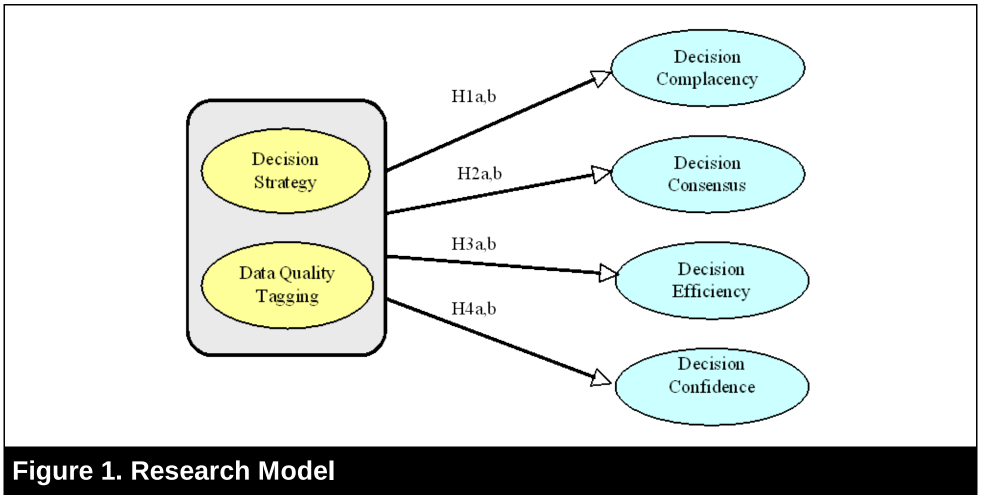


Figure 1. Research Model

In the additive strategy, an overall desirability score is calculated for each alternative by summing the assigned desirability scores of its individual attribute values. We rank alternatives 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 (i.e., multi-level) sort. Alternatives are initially sorted based on individual desirability scores of the attribute most important to the decision maker (i.e., 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. The implication is that a change in preferred decision choice with DQ tags is a consequence of a change in the priority given different attributes with respect to their relative importance as decision criteria. Further distinguishing the current research method from those used in earlier studies, we test this assumption by measuring and directly comparing whether DQ tags changed the relative priority assigned to the attribute tagged experimentally as being of the lowest quality (i.e., the poor quality attribute). In effect, this serves to verify any experimental findings with respect to complacency based on preferred decision choice. 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 be supported best 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:

> ***H1a:*** *decision makers are complacent with or without DQ tags for the additive strategy, and (H1b) decision makers are complacent with or without DQ tags for the EBA strategy.*
>
> ***H2a:*** *there is no difference in decision consensus with or without DQ tags for the additive strategy, and (H2b) there is no difference in decision consensus with or without DQ tags for the EBA strategy.*
>
> ***H3a:*** *there is no difference in decision efficiency with or without DQ tags for the additive strategy, and (H3b) there is no difference in decision efficiency with or without DQ tags for the EBA strategy.*
>
> ***H4a:*** *there is no difference in decision confidence with, as compared to without, DQ tags for the additive strategy, and (H4b) there is no difference in decision confidence with, as compared to without, DQ tags for the EBA strategy.*

The statistical analysis used is in accordance with standard statistical practice and recommendations (e.g., Pallant, 2001) and previous DQ tagging studies. Thus, we use a chi-squared statistic to test H1 and H2 based on a change in preferred decision choice, since they involve a categorical dependent variable. All the other hypotheses involve a continuous dependent variable. Therefore, we use either an independent samples t-test or a Mann-Whitney test, respectively, depending on whether the data is normally distributed or not.

Given prior research evidence of increased DQ tag usage with more experienced decision makers (see the Introductory section) and considering available resources, we used as participants those university students most likely to have decision-making and professional experience (i.e., postgraduate students enrolled in a masters or PhD degree program rather than undergraduates). Of the 62 participants in this study, 25 (i.e., 44 percent) had prior work experience and 10 (i.e., 16 percent) 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 (closely monitored by the investigators) and then nominate a confidence level using a five-point Likert scale ranging from very low to very high, rank the attributes in terms of their relative importance for the experimental decision task, and briefly explain their decision on answer sheets provided. As part of this explanation, participants were asked to write down on the answer sheet whether there were any attributes they ignored in their search and—if so—why.

In the context of the decision-making task, the hypotheses H1 through H4 are operationalized as follows:

H1: *There is no significant difference either in (i) the preferred apartment or in (ii) the ranking of the poor quality attribute with respect to its importance for the experimental decision task with as compared to without DQ tags.*

H2: *There is no significant difference in the proportion of decision makers selecting the preferred apartment with, as compared to without, DQ tags.*

H3: *There is no significant difference in the decision time with, as compared to without, DQ tags.*

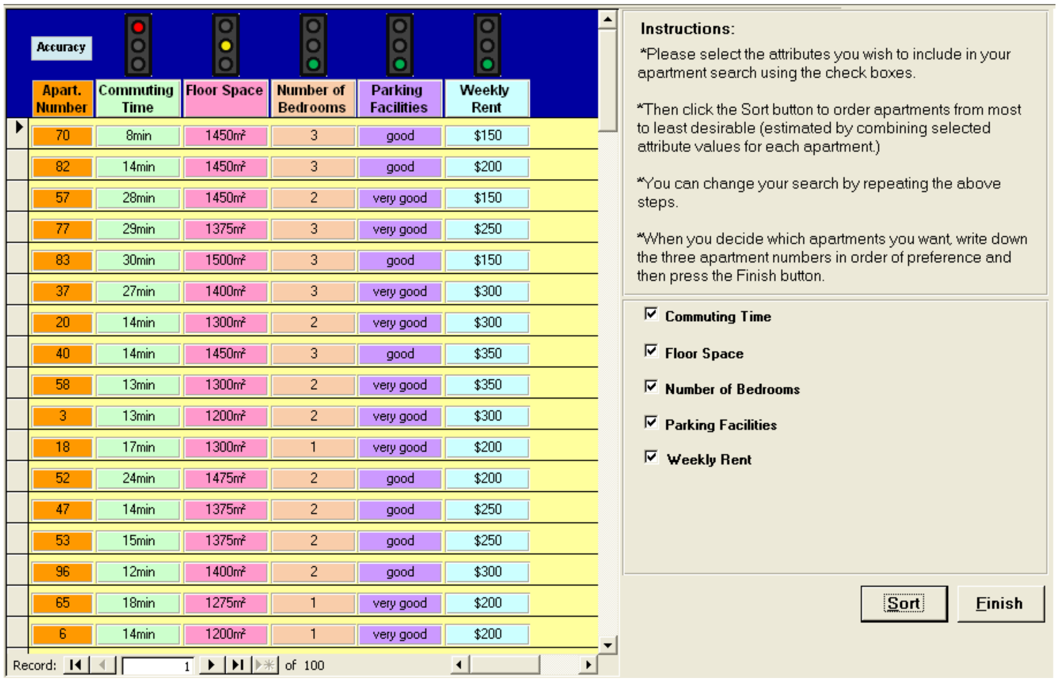
H4: *There is no significant difference in the nominated level of decision confidence with, as compared to without, DQ tags.*

We developed an interface, a set of instructions, and an answer sheet for each of the four different experimental treatments described previously. As in previous studies, the database alternatives were designed so that one apartment is clearly the most desirable without DQ information but is less desirable when DQ information is considered. In common with prior DQ tagging studies (to facilitate comparison of results), the rental property selection decision task is used and the attribute commuting time is tagged as having the lowest DQ value (i.e., is selected as the poor quality attribute) for those experimental interfaces with DQ tags. The experimental credibility of such tags is supported by comments both from managers of university housing databases (interviewed for the usability study) and from university students participating in our research work. Managers noted that students routinely notify them of data discrepancies. Written and verbal comments by student participants indicate that they are well aware that such rental property applications frequently contain errors, both in general and specifically with respect to commuting time estimates. For example, one participant noted specifically that—in common with the experiment—“travel time and distance were usually unreliable in web-based rental property applications that [she]…had used”.

In general, the experimental materials and DQ tag design we used differ significantly from earlier experimental designs (Chengular-Smith & Pazer, 1999; Fisher et al., 2003; Shanks & Tansley, 2002) in that they incorporate all of the recommendations resulting from the Usability Study described in the previous section. Furthermore, a unique aspect of this research design as opposed to that of earlier studies was that participants were asked to rank attributes by their relative importance to the decision task, permitting a direct measurement of whether the relative priority given to different attributes changed with DQ tags.

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.

We initially piloted the experimental design individually, with five postgraduate students and one 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 to be clear and did not result in any new suggestions, indicative of saturation.



Accuracy	🔴	🟡	🟢	🟢	🟢
Apart. Number	Commuting Time	Floor Space	Number of Bedrooms	Parking Facilities	Weekly Rent
70	8min	1450m^2	3	good	$150
82	14min	1450m^2	3	good	$200
57	28min	1450m^2	2	very good	$150
77	29min	1375m^2	3	very good	$250
83	30min	1500m^2	3	good	$150
37	27min	1400m^2	3	very good	$300
20	14min	1300m^2	2	very good	$300
40	14min	1450m^2	3	good	$350
58	13min	1300m^2	2	very good	$350
3	13min	1200m^2	2	very good	$300
18	17min	1300m^2	1	very good	$200
52	24min	1475m^2	2	good	$200
47	14min	1375m^2	2	good	$250
53	15min	1375m^2	2	good	$250
96	12min	1400m^2	2	good	$300
65	18min	1275m^2	1	very good	$200
6	14min	1200m^2	1	very good	$200

Record: 1 of 100

Instructions:

*Please select the attributes you wish to include in your apartment search using the check boxes.

*Then click the Sort button to order apartments from most to least desirable (estimated by combining selected attribute values for each apartment)

*You can change your search by repeating the above steps.

*When you decide which apartments you want, write down the three apartment numbers in order of preference and then press the Finish button.

☑ Commuting Time

☑ Floor Space

☑ Number of Bedrooms

☑ Parking Facilities

☑ Weekly Rent

Figure 2. Interface for Additive Decision-making Strategy with DQ Tags (Sorted)

We assigned participants to one of the four treatment groups 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 for the given 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. This process can be repeated with different attribute selections and/or rankings until the participant is satisfied and ready to make his or her apartment selection. For the treatments involving DQ tags, we checked completed answer sheets before participants left the laboratory to see if they used the poor quality attribute (commuting time). If so, we queried the participant 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. DQ Tagging Experiment: 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 (Pallant, 2001). 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 described in the previous section can be rejected.

Table 1. Analysis of Complacency (based on decision choice) and Consensus

	Decision Strategy	
	Additive	Elimination by Attributes
Complacency	$\chi^2 = 1.874$	$\chi^2 = .212$
	p = .171 (H1a)	p = .645 (H1b)
Consensus	$\chi^2 = 1.874$	$\chi^2 = .212$
	p = .171 (H2a)	p = .645 (H2b)

Table 2 verifies the finding with respect to complacency, since there is no significant difference shown in the priority ranking given to the attribute designated as being of the lowest quality (i.e., commuting time) for either decision strategy. Since significant variations from the normal distribution were evident for each decision strategy (based on a visual inspection of relevant histograms, the Kolmogorov-Smirnov test, and the Shapiro-Wilks test), we used the non-parametric Mann-Whitney test for data analysis. Thus, the mean rank and level of significance are shown with the mean and standard deviation for each treatment group.

Table 2. Analysis of Complacency (using the ranking of the poor quality attribute)

		Decision Strategy					
		Additive			Elimination by Attributes		
		Mean rank	Mean	SD	Mean rank	Mean	SD
Complacency	No Tags	17.73	3.07	1.44	14.00	2.07	.83
	Tags	14.38	2.56	1.36	17.65	2.76	1.52
		p = .286 (H1a**)**			p = .250 (H1b)		

Table 3 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 3 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.”

Table 3. Number of Participants Selecting Preferred and Other Apartments

		Number of Participants (% of participants selecting apartment from the set of apartments listed)	
		No Tags	Tags
Additive	Preferred	12 (80% for apt 70)	8 (50% for apt 70)
	Other	3 (20% for apt 77,83 or 98)	8 (50% for apt 5,33,37,47,49,62,77,83 or 98)
	Alternate	1 each (7% each for apt 77, 83 and 98)	2 (12% for apt 33)
	Total	15	16
EBA	Preferred	7 (50% for apt 5)	6 (35% for apt 5)
	Other	7 (50% for apt 33 or 66)	11 (65% for apt 16,33,38,44,66,98 or 100)
	Alternate	4 (29% for apt 33)	3 (17% for apt 66)
	Total	14	17

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 3 shows a decreased percentage (30 percent less for additive and 15 percent 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 between the desirability of the apartment preferred with tags and that preferred without tags.

Based on participant responses to the exit query described in the previous section, only 4 (1 using the EBA and 3 using the Additive decision-strategy) of the 33 participants given DQ information ignored the poor quality attribute (i.e., commuting time). Of those, 2 didn’t care about commuting time and 2 ignored it because of its poor 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 (i.e., share the intended interpretation of tag semantics). This suggests that the revised DQ tag design resulting from the recommendations of the earlier usability study was, in fact, understandable. All the others said they considered commuting time to be so important that they included it despite its poor 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 behavior by making assumptions about the degree of error ("even if commuting time is wrong, it is probably not too far off").

As described earlier for the analysis of complacency based on the priority ranking given the poor quality attribute (i.e., commuting time), we used the non-parametric Mann-Whitney test to analyze time and confidence, since they also showed deviations from normal distribution for each decision strategy. Table 4 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 4. Analysis of Time and Confidence

		Decision Strategy					
		Additive			Elimination by Attributes		
		Mean rank	Mean	SD	Mean rank	Mean	SD
Time	No Tags	12.67	4.14	1.92	14.18	5.29	2.81
	Tags	19.13	6.38	2.60	17.50	6.47	3.28
		p = .046* (H3a)			p = .308 (H3b)		
Confidence	No Tags	14.40	2.00	.76	14.29	2.00	.56
	Tags	17.50	2.31	.80	17.41	2.24	.56
		p = .305 (H4a)			p = .247 (H4b)		

The only significant change evident in decision outcomes when DQ tags were made available is an increase in the time required to make the decision for those treatments involving an additive rather than EBA decision strategy. This differs from earlier research (Chengular-Smith & Pazer, 1999; Fisher et al., 2003; Shanks & Tansley, 2002) that reported a significant change in preferred decision choice in conjunction with consensus in certain circumstances (and no change in consensus otherwise). However, the significantly increased decision time observed for DQ tags with the additive decision strategy, without any concomitant change in decision choice, is in agreement with Shanks and Tansley (2002). They suggest that more time may be required for the additive decision strategy because the impact of an individual attribute on the sort sequence is less obvious given the compensatory nature of the strategy. Decision makers may, therefore, find it more difficult to understand how the sort sequence would be affected if the attribute tagged as being of poor quality is not included in the sorting criteria.

In common with Shanks and Tansley (2002), the experiment 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 next section describes a protocol analysis study conducted to address this question.

6. Protocol Analysis Study: Research Method

Cognitive process tracing is a recognized data collection technique used in cognitive psychology and information systems research (e.g., Kim & Maher, 2008). In particular, protocol analysis (Ericsson & Simon, 1993) has been used for this purpose. This is a descriptive and interpretative technique that involves having participants verbalize their thoughts out loud as they complete some task, for example, decision making in the current context. This allows investigators to access decision makers' thought processes. Protocol analysis can, thus, be used as a means of comparing the thought processes of decision makers with and without tags, in order to better understand why DQ tags can affect decision time even when they do not impact the choice of which criteria to consider in the decision making or the final decision choice made.

We used concurrent protocol analysis to collect data about the cognitive processes of participants making the rental property selection decision from the DQ tagging experiment described above. Having participants verbalize their thoughts at the same time as the problem solving (concurrent protocol) rather than recalling their thoughts afterward (retrospective protocol) is recommended because it avoids the possibility of inaccurate memory recall (Ericsson & Simon, 1993, p. xiii) and allows correlation of thought processes with observable actions and information perceptually available to the participant (e.g., current state of an online interface or mouse position).

This verbal protocol technique is based on the assumption that in the course of solving a problem, people consciously construct a representation of the problem and the strategies used. Furthermore, a distinction is made between being asked to "think aloud" as compared to being asked to "explain," "describe," or "justify" what they are doing. There is evidence (Ericsson & Simon 1993, Preface) that the former does not change the nature of the problem solving process or the sequence of thought (except that more time may be required for verbalization as compared to silently performing the task), whereas the latter three activities require additional cognitive processes and thinking. Thus, when asked to think aloud, people "simply verbalize the information they attend to while generating the answer" (Ericsson & Simon, 1993, p. xiii). People are more accustomed to explanatory or descriptive verbalizations than thinking aloud, especially when there is another person present. Therefore, there are specific practices recommended as part of the experimental procedure in order to ensure that experimental participants understand what is required. These include careful wording of instructions (e.g., "think aloud whatever you say to yourself as you make the decision," "talk aloud constantly," or "act as if you are alone in the room"), seating the investigator out of the participants' sight (e.g., behind the participant), using practice tasks to accustom participants to verbalizing their thoughts, and giving reminders to "keep talking" or "think aloud" if the participant is silent.

Other types of observation can be used in conjunction with verbalizations to understand cognitive processes. Ericsson and Simon (1993, p. xv, 172-174) discuss observation of participant eye movements and information available to the participant's perceptions (e.g., current state of an online interface). Kim and Maher (2008) record the physical actions of problem solvers in addition to their verbalizations. Thus, a protocol consists of "the recorded behavior of the problem solver" (Kim & Maher 2008, p. 115) and the subsequent analysis takes into account both verbalizations and actions. We adopt this approach in the current study.

As described in the introduction, our goal is to understand participants' cognitive behavior in those circumstances when there was a significant difference in decision outcome observed. Therefore, only the additive decision strategy is considered in the protocol analysis study. We use protocol analysis to study the decision process in order to help explain why decision time was significantly impacted by the presence of DQ tags for this decision strategy. This means there were two different treatments in the protocol analysis, either with or without tags. The experimental materials used are the same as those used for the two additive treatments in the DQ tagging experiment described in the previous section, except that the answer sheet did not include information about start or finish times.

To ensure that participants think aloud rather than communicate, our research procedure follows all of the recommended practices described previously. Participants are given practice tasks—first paper-based and then online—until they are comfortable with thinking aloud. The practice tasks involve simple multi-criteria decisions such as making a menu selection or choosing a book to buy from a list of available titles with brief descriptions. Participants are then trained in the rental property selection interface from the DQ tagging experiment described in the previous section. They are asked to verbalize their thoughts while selecting a rental property using the online interface. The session is recorded using software called Morae with a video camera attached to the computer (facing the participant). This records and clocks both participant behavior (actions and verbalizations) and the online screen changes. The participant is then asked to fill out the answer sheet with demographic information. We conducted two pilot tests of the research procedure before collecting data for analysis.

Analysis of the data collected follows the recommendations in Ericsson and Simon (1993, Ch. 6, 7). This involves first analyzing the task and the protocols (i.e., recorded sessions) to define a coding scheme, i.e., the set of behaviors—both high and low-level—relevant to the task in question. Each protocol is then segmented and encoded based on the coding scheme, where each segment corresponds to one of the low-level behaviors defined in the coding scheme. Segments are then aggregated based on high-level behaviors, each of which is considered a separate behavior category. Individual protocols are analyzed independently by two coders and differences reconciled.

In order to compare the cognitive processes of decision makers with and without tags, we analyze the high-level behavior categories in three ways using three different graphs. The first analysis gives the average time proportion (i.e., percentage) spent in each high-level cognitive behavior category. This comparison shows in which category the main differences in cognitive behavior occurred with DQ tags. In order to understand which cognitive behaviors dominated during different stages of the decision-making task, the second analysis illustrates the average time proportion spent in each high-level cognitive behavior category for each of three equal time intervals. Finally, the third analysis shows the pattern of transitions between cognitive behaviors. Each of these analyses represents an average for all of the participants (i.e., across all protocol recordings) in one of the two treatment groups: with tags or without tags.

For each participant, the percentage of time spent in each category is calculated by dividing the actual time spent in that category by the total time duration of either the entire protocol recording (for the first analysis) or a single time interval (for the second analysis) and multiplying the resulting proportion by 100. The beginning, middle, and end time intervals used in the second analysis are calculated for a given participant by dividing their protocol recording into three equal and sequential time segments. For the third analysis, the total number of directed transitions between each two categories is calculated for each protocol and then averaged across protocols in a given treatment group.

The 13 participants in the study had a similar background to those in the DQ tagging experiment described in the previous section. They were all currently enrolled in or had completed a postgraduate (Masters or PhD) degree. Ten had previously used an actual online system to look for a place to live: the other three had prior experience with other online decisions. Ten had at least one year of prior professional experience and seven had prior managerial experience. The age of those participants specifying their age ranged from 21 to 52 (one participant did not specify). Seven participants had DQ tags (i.e., a traffic light symbol was displayed for each attribute column in the decision-making interface to indicate whether that attribute was of poor, medium, or good quality) and six participants did not.

6.1. Defining a Coding Scheme

The coding scheme consists of three levels: an abstract level describing the conceptual phase of the decision-making task, a middle level signifying underlying intentions, and a concrete level of directly observable behaviors. Ericsson and Simon (1993, Ch. 4, 5, 6) advocate that initial encoding be based on a concrete level with directly observable behaviors in the coding scheme. Encoding of protocols using only abstract level behaviors requires coders to infer from observed to abstract behaviors as they encode—potentially impacting the validity (i.e., mapping from the operational phenomenon to the correct theoretical construct) of the encoding process. However, it is difficult to generalize or conceptualize from the concrete level. Instead, Ericsson and Simon (1993, p. 273) recommend that aggregation of concrete -level behavior observations (e.g., based on strategies, solution phases) into abstract-level protocol segments be used for this purpose. This is the approach we follow.

We derived the abstract level of coding describing different categories of cognitive behavior from the three stages of problem solving defined by Dewey (1910) and the related three phases of decision making defined in Simon's (1977) decision-making model,[2] summarized as follows (listing first the problem-solving stage and then the corresponding decision-making phase in italics):

- *What is the problem? Intelligence (identifying and defining the decision task).*
- *What are the alternatives? Design (designing and analyzing the consequences of possible decision solutions, i.e., alternative decision choices)*
- *Which alternative is best? Choice (comparing the possible decision solutions to find the best solution)*

Although one would generally expect that a problem must be identified before possible solutions can be considered and that a final solution can be chosen only after it is articulated as a possible solution, the ordering of these phases within a given decision-making process may not be strictly sequential. Instead, these phases are interleaved, since a decision-making process typically involves solving a number of individual sub-decisions—each described by these same three processing phases. According to Simon, most of the time is normally spent in the Design phase. In the coding scheme listed below, we modified the definitions of these three phases as appropriate for the context of the current study and the specific decision task of selecting a rental property.

We initially defined the concrete and middle levels based on a preliminary examination of the protocols. We trialed the effectiveness of this scheme for coding on the first two protocols. Significant differences between the two coders and difficulties experienced in applying the initial coding scheme led us to a refinement of the coding scheme. When we then analyzed these two protocols and subsequent protocols using the refined coding scheme given below, the differences between the coders were fairly minimal. Table 5 below gives individual behavior codes for abstract (numeric bullets), middle (alphabetic bullets), and concrete (no bullets) levels of coding. Each row of the table gives the behavior definition followed by the code in brackets. To illustrate, one participant commented on the consequences of a short commuting time for a specific on-screen apartment as follows: "The apartment has a commuting time of only eight minutes, so I can get a bike, use a bike." This is coded in the Search [SE] sub-category (a) of the Design [DES] category 2 as an example of [EvalApt] behavior.

[2] In common with Simon, we focus on the first three phases of decision making and, thus, do not discuss decision implementation or review.

Table 5. Coding Scheme for the Decision Task of Selecting a Rental Property

Abbreviations Used: “apt.” or “apts.” for apartment(s); “apt#.” for apartment number
1. “Intelligence”: identify/define problem [INT]
(a) Identify problem: clarification of interface or decision-task [CL]
Read or reference on-screen or written instructions *[Instr]*
Ask question about on-line interface or decision task *[Ques]*
Reflect on nature of interface *[RInterface]*
Reflect on nature of decision task, could include specifying contextual assumptions, eg. single or with family, own car or bicycle *[RTask]*
(b) *Define problem: specify or re-specify sort [SP]*
Move mouse over check box *[MouseCB]*
Select or de-select check box *[SelectCB]*
Define verbally attributes of interest while specifying sort *[DefAttrib]*
Discuss relative importance or preferred values of attributes in general (not based on specific apts.) *[Attr]*
Discuss selection or non-selection of an attribute for the sort based on DQ tag values *[Tag]*
Press sort button *[Sort]*
2.“Design”: search for and select possible solutions (ie. alternatives) [DES]
(a) Search for possible solutions: look for possible apts. and clarify criteria required for solutions [SE]
Look at on-screen apts *[LookApt]*
Evaluate attribute values of on-screen apt. against preferred values or in terms of consequences *[EvalApt]*
Compare multiple on-screen apts. with respect to preferred attribute values *[CompApt]*
Search through on-screen apts. to find those with preferred attribute values *[FindPreferApt]*
Scroll through screens to find more apts. *[Scroll]*
Compare on-screen apts. to previously selected apts. *[CompAptToSel]*
Clarify criteria required for solution *[ClarifyCrit]*
(b) Select possible solutions: select a specific apt. [SL]
Verbally note apt as being of interest as a possible solution *[SelAptV]*
Write down an apt. number on paper, indicating that it is of interest as a possible solution *[SelApt]*
Clarify why apt. is of interest (ie. in terms of having preferred attribute values or the consequences of having certain attribute values) while (before, during, just after) selecting apt. *[ClarifyAptSel]*
3. “Choice”: Evaluate alternative solutions and make final choice of solution [CHO]
(a) Evaluate alternative solutions: evaluate or compare previously selected apts. [EV]
Search through on-screen apts. to search for a previously selected apt. by its apt#.*[FindSel]*
Consider how attribute values (written down and/or on-screen) of selected apt. match preferences *[EvalSel]*
Compare selected apts. to each other, usually in terms of how attribute values match preferences *[CompSel]*
Order selected apts. by preference (either verbally or in written form) *[OrderSel]*
(b) Make final choice of solution: select preferred apt. [DE]
Verbalize or write down preferred apt. *[FinalCho]*
Clarify why apt. is preferred *[ClarifyFinalCho]*

7. Protocol Analysis Study: Results

In common with the DQ tagging experiment described earlier in this paper, the majority of the participants (12 out of 13) considered commuting time in their decision making regardless of whether DQ tags were included or not. Six out of the seven participants given DQ tags used commuting time even though they understood that it was tagged as being of very poor quality. Similar justifications were given for ignoring the DQ tag information associated with commuting time (i.e., when deciding which decision criteria to consider) with respect to the importance of commuting time for this decision and the assumption that—if wrong—the commuting time would probably not be too inaccurate. Another participant noted that “you couldn’t really be sure whether an apartment was suitable until you lived there” (with respect to assessing traffic noise, neighbors, etc.) and that other important information was missing (e.g., local crime rate and conveniences such as access to public transport), so “the decision was just a best guess” regardless of whether the commuting time given is correct. Interestingly, during the course of the decision-making session, two participants repeatedly commented verbally that they should not be considering commuting time because it was not reliable but ultimately used it anyway.

The three different analyses described in the previous section are shown in Figures 3, 4, and 5, respectively.

We can see from Figure 3 that the results of either treatment are consistent with the assertion in Simon (1977) that decision makers spend the greatest proportion of their time in the Design phase. There is, however, a noticeable difference in the proportion of time spent in the Intelligence phase between the two treatments. Intelligence comprises almost one-third of the total time used for decision making with DQ tags but less than one-sixth of the total time used without DQ tags. Instead, an increasing proportion of the time is spent in the other two decision-making phases—especially in the Design phase—for the treatment without DQ tags. This suggests that the presence of DQ tags required that more effort be focused on problem identification and definition.

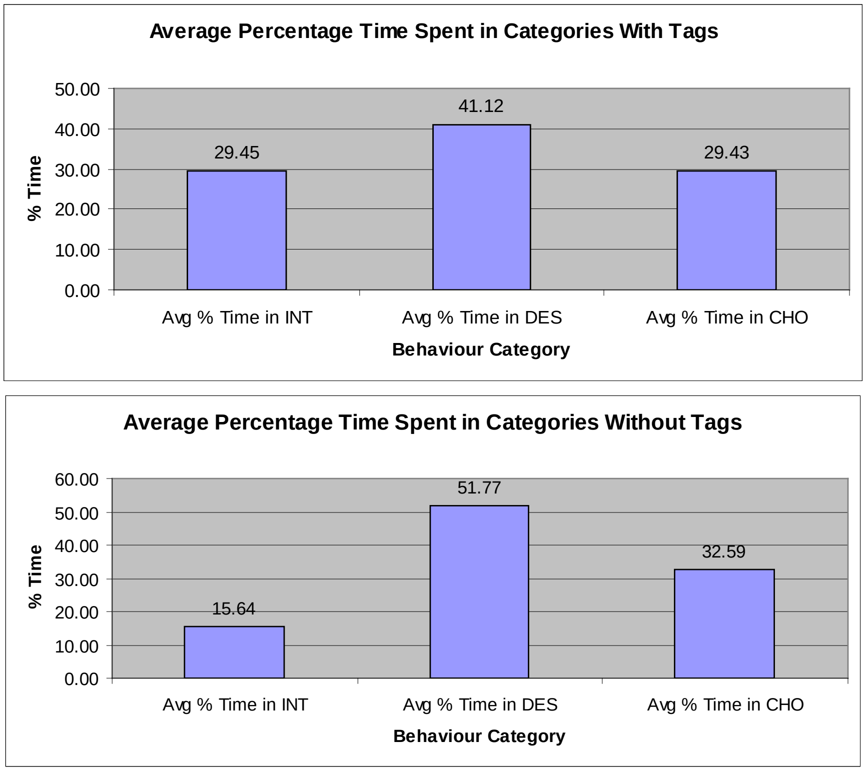


Figure 3. Percentage Time Spent in Intelligence (INT), Design (DES), and Choice (CHO) Behavior Categories

In Figure 4, the shape of the plot for each cognitive behavior category is similar, irrespective of treatment (i.e., with or without DQ tags). For example, Figure 4 shows that—for either treatment—the proportion of time spent in Intelligence category behavior peaks in the first (beginning) time interval, with much less occurring in the second (middle) time interval, and the least evident in the third (end) time interval. In contrast, the plot of Choice behavior is strictly increasing from beginning to end time intervals, and Design behavior peaks in the middle time interval and is lowest in the end interval for both treatments.

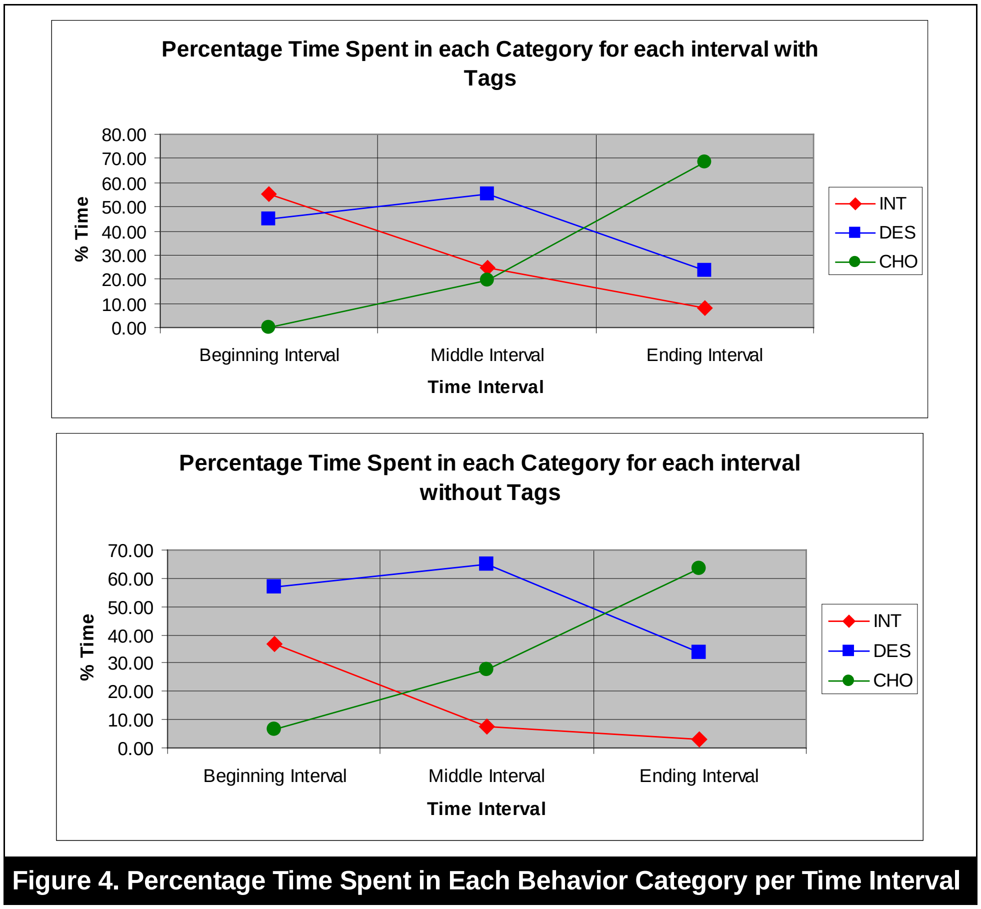


Figure 4. Percentage Time Spent in Each Behavior Category per Time Interval

The relative dominance of the different categories of behavior in the first time interval is, however, quite different between the two treatments. Whereas Intelligence dominates the beginning time interval with DQ tags, Design dominates the beginning time interval without DQ tags. Furthermore, although in the middle interval Design behavior predominates regardless of treatment, Intelligence behavior is much more marked with, as compared to without, DQ tags. With DQ tags, Intelligence comprises one-fourth of the middle time interval, whereas only one-fifth of the time is spent on Choice. Without DQ tags, Intelligence uses less than one-tenth of the middle time interval, whereas almost one-third of the time is spent in Choice.

When we consider Figure 5, we see that the two treatments each have a different transition pattern among the three cognitive behavior categories, especially with respect to transitions occurring either between Intelligence and Design or between Design and Choice. Transitions between Intelligence and Design predominate in the treatment with DQ tags (eight times as many as the transitions between Design and Choice), whereas they are evenly balanced in number between Design and Choice in the treatment without DQ tags. This indicates that there is more iteration in the earlier phases of decision making when DQ tags are present.

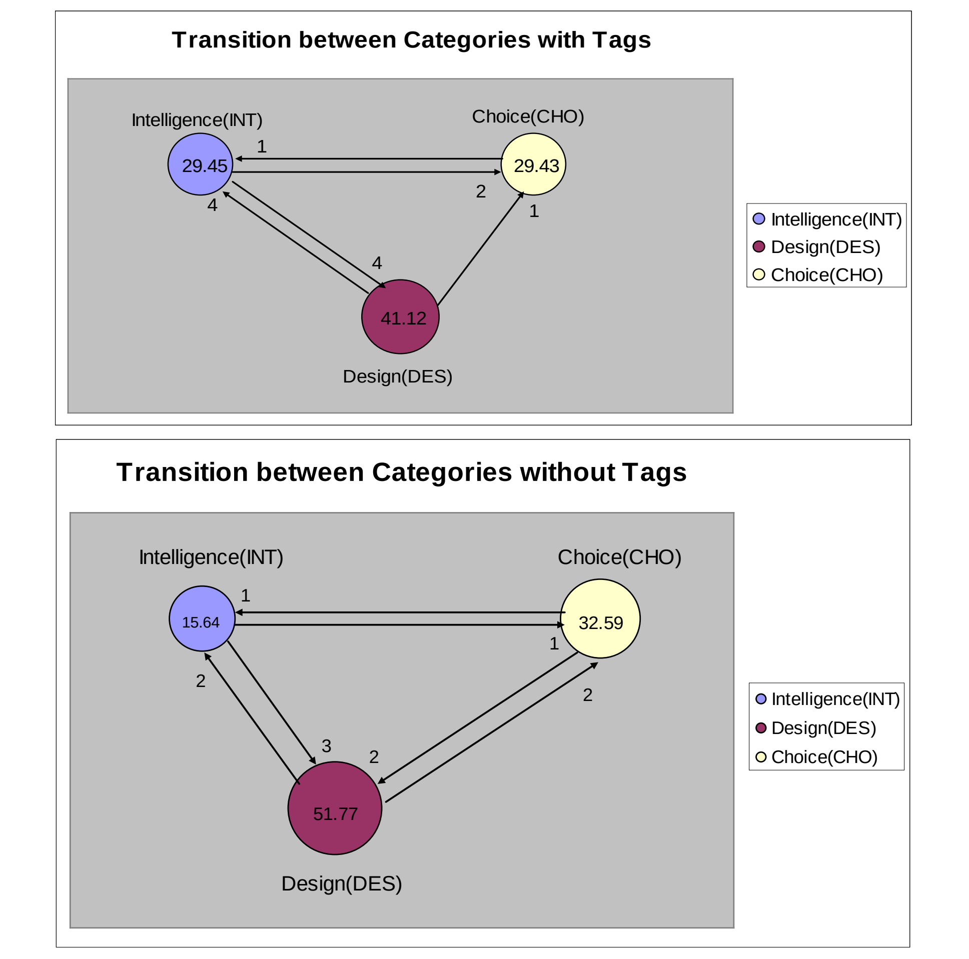


Figure 5. Total Number of Transitions between Behavior Categories

In combination, the analyses suggest that DQ tags impact the nature of the cognitive processes used in decision making even when they do not have a significant impact on decision choice. The results of this study suggest that the presence of DQ tags requires an increase in the time (both in terms of overall total and proportion compared to other categories of decision-making behavior) required to identify and define the problem, especially during the first two-thirds of the decision-making process. Thus, Intelligence behavior is quite prominent throughout the first two time intervals of the decision-making process with DQ tags, but exhibits a marked fall-off after the first time interval in the decision-making process without DQ tags. The changed pattern of transitions with, as compared to without, DQ tags suggests that the process of generating possible solutions is increasingly interrupted by the need for further problem clarification, associated with a delay in the choice of final solution.

8. Discussion

The current DQ tagging experiment is distinguished from earlier quantitative experiments (Chengular-Smith & Pazer, 1999; Fisher et al., 2003; Shanks & Tansley, 2002) 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 the above mentioned DQ tagging work by confirming the results of one of the studies, the current study raises more questions. Other than changes in DQ tag design for semantics and usability, we made experimental design choices to ensure consistency with earlier research designs and to focus on those decision contexts consistently reported as being associated with the highest levels of DQ tag usage in earlier studies. Nevertheless, the current study differed significantly from earlier DQ tagging research in that there was no evidence that DQ tags significantly changed the preferred decision choice, but there was still some indication (based on Table 3) of an overall decline in consensus. In contrast, the earlier three studies reported specific circumstances where DQ tags were associated with a change in preferred decision choice, and changes in consensus were only observed in conjunction with such a change. We consider two possible explanations for the difference in experimental outcomes: the change in DQ tag design in the current study 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 may have contributed to the observed difference in results. In particular, it is not surprising that there may be a difference in research findings when issues that could have potentially impacted the degree of experimental soundness in earlier work, such as the understandability of the DQ tags used, are explicitly addressed in experimental design. However, if a more usable experimental DQ tag design resulted in DQ tags that were more consistently interpreted or relevant; it seems intuitively more likely that there should be more use of DQ tags. In fact, the opposite was true.

Participant comments suggest that the choice of attribute tagged as being of poor quality may have influenced the observed outcome in the current experiment. Even when available DQ information indicated that commuting time was of poor quality, feedback from participants reveals that they considered the attribute much too important to ignore due to concerns about petrol prices, the environment, and traffic. The relevance of these comments to the observed results is reinforced by the fact that there was no evidence of a significant shift in the relative priority ranking participants gave to commuting time. Given that there has been a dramatic increase in petrol prices, environmental awareness, and traffic in the years since the three earlier studies were conducted, there may be a greater motivation to include commuting time than was the case in these earlier studies. Thus, it would be worthwhile to consider the effect of changing the attribute selected as being of the lowest quality (i.e., the poor quality attribute) on DQ tag use. If the attribute selected to be tagged as being of poor quality were one that participants generally viewed as useful but not critical for the decision, then those participants given DQ tags might be more likely to ignore this attribute and those participants without DQ tags to use it.

It is important to note, however, that none of the DQ tagging experiments reported in the literature to date provide unqualified support for DQ tags. For example, a decrease in decision consensus was consistently associated with observations of changed decision choice with tags. Furthermore, evidence of DQ tag use was limited to specific treatments. The current experiment addressing DQ tag use in online decision making can be most meaningfully compared to Shanks and Tansley (2002), since Chengular-Smith and Pazer (1999) and Fisher et al. (2003) were paper-based and involved fewer than 10 alternatives. DQ tags were associated with changed decision choice in only one of four comparisons in Shanks and Tansley (2002), using a very simple decision task and only four attributes (i.e., potential decision criteria). Many online decisions would be more complex than this; therefore, empirical evidence thus far would seem to indicate that DQ tags would not be generally useful. However, a limitation common to both the current and previous DQ tagging studies is the artificial and simplistic nature of the decision task used in a laboratory setting.

One of the participants in the DQ tag usability study conducted earlier by the authors (Price & Shanks, 2009) posited that decision makers were most likely to be influenced by a DQ tag if it indicated that an attribute of moderate importance to the decision task was of poor quality. In other words, attributes of critical importance would be used and those of very little importance ignored, regardless of their associated DQ tag value. Such a view may be especially valid given the above mentioned limitations of the decision tasks used in DQ tagging research to date. It may be difficult to find an attribute that is regarded as moderately important to the majority of participants for an artificial decision task involving only a small set of attributes that can be considered as decision criteria. If true, this explanation for the negative results of the current experiment would reinforce the view that the current evidence does not support the general use of DQ tags, since they would impact at most only consideration of those attributes moderately important to the decision task at hand. However, one other factor that should be considered is the non-critical nature of the decision tasks considered in DQ tagging research to date, as illustrated by the rental property selection task used in the current experiment (chosen to facilitate comparison with previous research).

Since a poor choice of rental property selection is not immediately obvious and does not ordinarily have serious consequences, laboratory subjects are not overly concerned about whether the use of poor quality data will negatively impact the decision made. The participant comments described in the experimental and protocol analysis result sections are congruent with this view. It may well be that quite different behavior may be observed when complex and critical decision tasks are considered, especially when examined in a realistic context using situated research techniques such as case studies. If the impact of basing a decision on an attribute known to be of poor quality is both immediately obvious and/or catastrophic (as, for instance, in an emergency room or disaster relief situation), then there is potentially more motivation for decision makers to use DQ tags even if it complicates decision making. In some cases, such tags may be useful for prioritizing information sources in decision making. For example, DQ information may be used in clinical diagnoses to distinguish between the relative reliability of different testing techniques in a hospital setting in order to more effectively weight the importance of test results.

In line with the findings reported by Shanks and Tansley (2002), the DQ tagging experiment described in this paper further shows that DQ tags can significantly decrease decision efficiency even when the preferred decision choice is not changed. The fact that these outcomes were associated in both studies with a less transparent (i.e., easily understandable) decision strategy is consistent with the general agreement in laboratory-based DQ tagging research to date that DQ use decreases as cognitive load increases.

This behavior can be explained in terms of the impact DQ tags have on decision process, as revealed by the cognitive process tracing study described in this paper. As evident from Figures 3, 4, and 5, respectively, the presence of DQ tags increases the overall proportion of time spent in the intelligence phase of decision making, the number of sequential time intervals in which intelligence behavior plays a major role in decision making, and the iteration between intelligence and design activities. Decision makers apparently devote considerable effort thinking about how to use the DQ tag information provided when defining the problem (intelligence behavior), and this complicates the design phase of decision making, potentially increasing decision time. However, they ultimately give the DQ information lower priority than other factors (e.g., the perceived importance of commuting time to the decision) and, thus, disregard it when selecting decision criteria.

As explained earlier, it is important to note that the cognitive processing pattern observed with DQ tags may be a function of the decision-making domains and laboratory-based research techniques used to date. The same participant in the usability study discussed earlier suggested further that use of DQ tags would require too much effort for infrequent users of an application (see Price & Shanks, 2009), and earlier research by Fisher (2003) suggests that the same is true for novice (as opposed to expert) users. Consideration and use of DQ tags may involve considerably less cognitive load for expert decision makers, since they are likely to be familiar with and frequent users of a given decision task and domain. Consistent with evidence from the quantitative DQ tagging experiment, participant comments during the cognitive process tracing study reveal that the potential impact of using poor quality data on decision effectiveness is not considered that significant. The nature of the application domain and the research technique employed are, thus, likely to be major factors influencing the degree to which DQ tags are used by decision makers and their effect on cognitive process.

9. Conclusion

We consider, in turn, the contributions of the DQ tagging experiment, the contributions of the cognitive process tracing study, recommendations for future research, and the implications for the adoption of DQ tags in practice.

The DQ tagging experiment examining the impact of DQ tags on decision outcomes (1) explicitly addresses semantic and usability issues in DQ tag design not considered in earlier research that could impact experimental soundness,[3] (2) contains an additional novel measure to verify and help understand the findings with respect to decision complacency (i.e., uses a dual measure of complacency indirectly based on decision choice as in previous experiments and directly based on changed prioritization of the poor quality attribute), and (3) focuses on areas of disagreement in earlier work with respect to the effect decision strategy has on DQ tag use. Thus, the DQ tagging experiment reported here represents an extension of earlier DQ tagging work, a research approach whose value to the investigative process is highlighted in Berthon et al. (2002) and is discussed with specific reference to DQ tagging research in the introductory section.

The current quantitative experiments confirmed previous findings by Shanks and Tansley (2002) that DQ tags can significantly increase decision time even when decision choice is not affected. However, the reported outcomes differ notably from that of all previous DQ tagging research in that there was no evidence that DQ tags changed decision choice, despite the fact that we chose an experimental design to facilitate comparison with previous work and to focus on those decision contexts consistently reported as being associated with the highest levels of DQ tag usage in earlier studies. Furthermore, there was some indication of reduced consensus with DQ tags even when neither the preferred decision choice nor ranking of decision criteria was impacted. It is clear that the extra financial and cognitive costs involved in maintaining and using DQ tags, respectively, cannot be justified based on such evidence.

Whereas DQ tagging research to date has considered only decision outcomes, the cognitive process tracing study described here satisfies the need—acknowledged in earlier DQ tagging research (Fisher et al. 2003)—for investigation of how DQ tags impact the actual decision-making process. The results showed that the generation of problem solutions is delayed by the increased time required for problem definition with DQ tags. This helps us to understand the observed impact of DQ tags on decision outcomes, thus explaining why decision time can be impacted by DQ tags even when other decision outcomes are not. Furthermore, the cognitive process tracing study serves to motivate and demonstrate the use of protocol analysis in the context of DQ tagging research.

The generalizability of this work should be tested in other application domains and using other research methods. In particular, future research should consider application domains such as hospital emergency rooms or disaster response involving critical decision tasks where the potential consequences of basing decisions on flawed data are immediately obvious and more serious. Decision makers may be willing to use DQ tags in such situations despite the additional cognitive load required; however, it is difficult to reproduce the same level of urgency in a laboratory context. Thus, case studies involving field research would be useful to address the limitations of a laboratory-based research approach with respect to scope and realism.

The implications of this work for practitioners are largely cautionary. Given the indications that the presence of DQ tags can decrease decision efficiency and consensus even when they do not influence decision choice, any implementation of DQ tagging in practice should be preceded by a careful investigation of its potential benefits and costs in the specific context considered. The results of the protocol analysis suggest that, where DQ tagging is adopted, organizations should consider strategies to minimize the extra cognitive load on decision makers, especially in the earlier stages of decision making. Such strategies could include selective deployment of DQ tags based on identification of specific contexts likely to have maximal benefit and minimal costs (e.g., for frequent rather than occasional users), training in effective use of DQ tags, and explicit consideration of user requirements and usability issues. The usability study described in this paper could potentially serve as a source of initial guidelines for DQ tag design and/or a precedent for the collection of specific user requirements using contextual inquiry techniques.

[3] The understandability of the resulting DQ tag design is demonstrated by the results of the exit query as discussed in the Experimental Results section, which indicated that all but one of the respondents understood the intended meaning of the DQ tags.

Acknowledgements

We thank the Australian Research Council for its funding; Ranjani Nagarajan, Ganesh A. and Rahul Sinha for their help in preparing experimental materials, conducting lab sessions, and data analysis; and Peter Benda for his invaluable advice and support throughout.

References

Benyon, P., Turner, P. & Turner, S. (2005). *Designing interactive systems.* Harlow: Addison-Wesley.

Berthon, P., Pitt, L., Ewing, M., & Carr, C. (2002). Potential research space in MIS: a framework for envisioning and evaluating research replication, extension, and generation. *Information Systems Research*, *13*(4), 416-427.

Beyer, H., & Holtzblatt, K. (1998). *Contextual design: defining customer-centered systems.* San Francisco: Morgan Kaufmann.

Chengalur-Smith, I.N., & Pazer, H.L. (1999). The impact of data quality information on decision making: an exploratory analysis. *IEEE Transactions in Knowledge and Data Engineering*, *11*(6), 853-864.

Dewey, J. (1910). *How we think*. New York: D.C. Heath & Company.

Eppler, M.J. (2001). The concept of information quality: an interdisciplinary evaluation of recent information quality frameworks. *Studies in Communication Sciences*, *1*, 167-182.

Even, A., Shankaranarayanan, G., & Watts, S. (2006). Enhancing decision making with process metadata: theoretical framework, research tool, and exploratory examination. *Proceedings 39th Hawaii International Conference on System Sciences (HICSS2006),* 1-10.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis.* Cambridge, Mass: The MIT Press.

Fisher, C., Chengular-Smith, I.N., & Ballou, D. (2003). The impact of experience and time on the use of data quality information in decision making. *Information Systems Research*, *14*(2), 170-188.

Holtzblatt, K., Wendell, J. & Wood, S. (2005). *Rapid contextual design: a how-to guide to key techniques for user-centered design*. San Francisco: Morgan Kaufmann.

Kim, M. J. & Maher, M.L. (2008). The impact of tangible user interfaces on designers' spatial cognition. *Human-Computer Interaction, 23*(2), 101-137.

Nelson, R., Todd, P., & Wixom, B. (2005). Antecedents of information and system quality: an empirical examination within the context of data warehousing. *Journal of Management Information Systems*, *21*(4), 199-235.

Neuman, W. L. (2006). *Social research methods: qualitative and quantitative approaches, 6th edition*. Boston: Allyn & Bacon.

Pallant, J. (2001). *SPSS survival manual a step by step guide to data analysis using SPSS for Windows (Version 10)*. Crows Nest, NSW, Australia: Allen & Unwin.

Payne, J., Bettman, J., & Johnson, E. (1993). *The adaptive decision maker*. Cambridge: Cambridge Univ. Press.

Price, R. and Shanks, G. (2009) Representing Data Quality Information Usably, Technical Report 2009/3, Clayton School of Information Technology, Monash University, 1-19.

Price, R. & Shanks, G. (2005). A semiotic information quality framework: development and comparative analysis, *Journal of Information Technology*, *20*(2), 88-102.

Shankaranarayanan, G., Even, A. & Watts, S. (2006). The role of process metadata and data quality perceptions in decision making: an empirical framework and investigation. *Journal of Information Technology Management*, *17*(1), 50-67.

Shankaranarayanan, G. & Cai, Y. (2006). Supporting data quality management in decision-making. *Decision Support Systems*, *42*(1), 302-217.

Shankaranarayanan, G., Ziad, M., & Wang, R. (2003). Managing data quality in dynamic decision environments: an information product approach, *Journal of Database Management*, *14*(4), 14-32.

Shanks, G. & Tansley, E. (2002). Data quality tagging and decision outcomes: an experimental study. *Proceedings IFIP Conference on Decision Making and Decision Support in the Internet Age*, Cork, 399-410.

Simon, H.A. (1977). *The new science of management decision.* Englewood Cliffs, N.J: Prentice-Hall, Inc.

About the Authors

Rosanne J. PRICE is an Adjunct Research Fellow in the School of Information Technology at Monash University, Australia. Her research interests and international journal and conference publications focus on data quality, databases, information systems (spatiotemporal and multimedia), and object-oriented and conceptual modeling. She received her PhD from Monash University. She has had over seventeen years of academic and professional experience, including research positions at Monash University (Senior Research Fellow) and the University of Melbourne (Australian Postdoctoral Fellow) and research/lecturing positions at University of Melbourne, RMIT, and European divisions of Boston University and the University of Maryland.

Graeme SHANKS is an Australian Professorial Fellow in the Department of Information Systems at the University of Melbourne. His research interests focus on business analytics, the implementation and impact of information systems, data quality and conceptual modeling. Graeme has published in journals including *MIS Quarterly, Journal of Information Technology, Information Systems Journal, Information & Management, Electronic Commerce Research, Journal of Strategic Information Systems, Information Systems, Behaviour and Information Technology, Communications of the AIS, Communications of the ACM, and Requirements Engineering.* He is a member of the editorial boards of six journals and was recently a member of the Australian Research Council College of Experts. Prior to becoming an academic, Graeme worked for a number of years as programmer, programmer-analyst and project leader in several large organizations.