Task-dependence of subjective believability in integration of scientific data

Ahmed Abuhalimeh*

IQ Programme,
University of Arkansas at Little Rock,
2801 South University Avenue,
Little Rock, AR 72204, USA
E-mail: aaabuhalime@ualr.edu
*Corresponding author

M. Eduard Tudoreanu

Information Science,
University of Arkansas at Little Rock,
2801 South University Avenue,
Little Rock, AR 72204, USA
E-mail: metudoreanu@ualr.edu

Thikra Mustafa

Nanotechnology Centre,
University of Arkansas at Little Rock,
2801 South University Avenue,
Little Rock, AR 72204, USA
E-mail: tamustafa@ualr.edu

Abstract: Believability is one of the major information quality dimensions that plays a role in the operational fitness and sound decision making. This paper presents an empirical evaluation of how people perceive believability of data shown through visual and textual representations. Integration of text and images is also studied with respect to believability. The subjective assessment exhibits variation for different types of data sources: textual, image, and both. The manner in which believability varies appears to be heavily dependent on task. Some tasks are more believable when text is integrated with images, others do not benefit from the combination. The results may be influenced by possible bias towards particular data. The data is the result of scientific research into the process of incubation of the bone cells with gold nanoparticles. This research was selected for our study because it alleviates the effect of the accuracy dimension on the assessment of believability. These results are complemented by previous studies on subjective perception of accuracy, and show a non-linear perception of information quality.

Keywords: believability; subjective quality; data quality; information quality; visualisation integration.

Reference to this paper should be made as follows: Abuhalimeh, A., Tudoreanu, M.E. and Mustafa, T. (2012) 'Task-dependence of subjective believability in integration of scientific data', *Int. J. Information Quality*, Vol. 3, No. 1, pp.71–84.

Biographical notes: Ahmed Abuhalimeh's research interests are related to information quality, information visualisations, human-computer applications/interaction, and medical and health data applications. His recent research is integrating information quality in visual analytics. His research focus on subjective IQ (SIQ) dimensions, which are dimensions that typically require a user's opinion and do not have a clear mathematical technique for finding their value. He received appointment at the American University in United Arab Emirates; he accepted the position of Assistant Professor in the Department of Computer Engineering. He is currently a Research Administrator at the University of Arkansas at Little Rock.

M. Eduard Tudoreanu's research interests are related to the use of graphics to better understand data and the processes that affect data. His work encompasses areas in information visualisation, human-computer interaction, information quality, and virtual reality. He is interested in the impact of advanced interaction techniques on the capacity of a user to gain insight into virtual environments that have a high density of graphical objects, often a result of rendering large, complex data sets. He has been the Keynote Speaker at ABSEL 2010, served as a Panellist for the National Science Foundation, and received grants from governmental and corporate sponsors.

Thikra Mustafa received her BS in Biology from University of Baghdad and MS from Al-Nahrain University Baghdad, Iraq. Currently, she is a PhD student in Applied Science at the Nanotechnology Center, Graduate Institute of Technology, University of Arkansas at Little Rock. Her areas of interests are focusing on synthesis and characterisation of nanomaterials, application of nanomaterials in bionanomedical field and cancer fighting, and studies the interaction of nanomaterial with animal cell (normal and cancer cells) using the electron microscope as visualisation tool.

1 Introduction

Believability, defined as the extent to which information is true and credible (Pipino et al., 2002; Wang and Strong, 1996), is one of the information quality (IQ) dimensions that can be best determined/assessed through subjective assessment of the information users rather than through algorithmic means. Some other quality dimensions, such as value-added or reputation, are also intrinsically dependent on the human actor, while others may become subjective in certain situations.

The term we introduce to describe quality measures that cannot be determined by a computer alone is subjective information quality or SIQ (Abuhalimeh et al., 2010). SIQ may not necessarily behave the same as the precisely computed measurements because they involve human factors and human psychology. Assessment of SIQ may be more application and situation specific, and rules for determining such quality may be different than statistical calculations. For example, people may find faults in data that is very true and credible, and may find the combination of two poor data sources to be more than the sum or average of the parts.

The aspects of SIQ covered in this paper include how subjective rating varies with different pieces of information displayed, and how additional information influences the assessment of believability. Our results are based on a study on the perceived believability of the concentration of gold particles added to the bone cells. The study employed data that can be easily judged by an average person. The data was obtained from the Nanotechnology Centre at University of Arkansas at Little Rock. The investigators introduced two types of data: lab notes (textual) and microscopy (image). The study assessed the believability of each type of data as well as the believability of the integration of text and image.

Believability was assessed with regard to two different concentrations of gold particles. Participants were asked to express their belief that the presented data was of a given concentration. The design of the study surreptitiously forced participants to provide their opinion based solely on belief because scientists could not determine any difference in results between the two concentrations. Thus, believability was measured alone without interference from other IQ dimensions, more importantly accuracy, which can skew the results for believability when people tend to not believe data that appears inaccurate. Note that choosing experts for the study, would result in them validating accuracy rather than believability because experts already use and trust these visual tools and lab notes.

The study revealed that believability varies with the type of data source, image, text, or both, and that it behaves differently for each task. Users assessed themselves as being neutral to confident in their results, with the text data source scoring the lowest, and the image scoring the highest.

The remainder of the paper is organised as follows: the next section discusses related work, followed by a description of study and the results. The paper concludes with a discussion and future work plans.

2 Related work

Pipino et al. (2002) present approaches that combine the subjective and objective assessments of data quality, however, their approaches do not ask for an estimation of the believability from participants. Their approaches are based on mathematical models, and focus on the data from one source. Our study provides visualisations techniques and aims to help in developing a method that will enable users to better estimate the quality of the data coming from different sources, due to the lack of the statistical methods for assessing SIQ.

Mo and Zheng (2008) present a method for measuring data quality in data integration. They focus on algorithmic methods for believability and, unlike our work, do not include the human's experiences with quality.

Nicolas and Madnick (2008) present the main concepts of a model for representing and storing data provenance, which includes an ontology of the sub-dimensions of data believability. They use aggregation operators to compute believability across the sub-dimensions of data believability. Our work focus on subjective evaluation of believability through visualisations techniques and aims to help in developing a method that will enable users to better estimate the quality of the data coming from different sources, and may be better at determining SIQ measures that the statistical methods employed for their data provenance model.

The research of Huerta and Ryan (2003) examines the factors affecting the credibility of online information. It uses the elaboration likelihood model (Petty and Cacioppo, 1986) as a theoretical framework, proposing a comprehensive model that includes factors from traditional means of communication and the Web. A field experiment was conducted that manipulated quality of content, reputation of the website owner, attractiveness, modality of exposure, and simulation. Out of these factors, quality of content and reputation of the website owner show statistical significance in the expected direction. Our study researches textual and visual data sources, and focuses on believability.

Lee et al. (2002) developed a methodology, called AIM quality (AIMQ), to form a basis for IQ assessment and benchmarking. The methodology is illustrated through its application to five major organisations. The methodology encompasses a model of IQ, a questionnaire to measure IQ, and analysis techniques for interpreting the IQ measures. They developed and validated the questionnaire and used it to collect data on the status of organisational IQ. These pieces of data are used to assess and benchmark IQ for four quadrants of the model, which rely on questionnaires to find IQ scores. Our study uses different data types (text and visualisations) from different sources and we aim to help in developing a method that will enable users/organisations to better estimate the believability of the data coming from different sources.

Bobrowski et al. (1999) presented a methodology to measure data quality within organisations. First, a list of IQ criteria must be set up. These IQ criteria are divided into directly and indirectly assessed criteria. Scores for the indirectly assessed IQ criteria are computed from the directly assessed IQ criteria. In order to assess the direct criteria, traditional software metrics techniques are applied. These techniques measure data quality following the goal-question-metric methodology: For each directly assessed criterion, a question is set up that characterises the criterion, and then a metric is derived to answer this question, giving a precise evaluation of the quality. From these metrics a user questionnaire is set up which is based on samples of the database rely on questionnaires to find IQ scores.

Both AIMQ and the approach of Bobrowski, Marre, and Yankelevich rely on questionnaires to find IQ scores. Our study examines people’s assessment of IQ based on their opinions and interaction with the information through different scenarios using samples of different data types (text, images, or both) in order to better understand how believability is influenced by data stemming from different sources and presented in a visual format.

Klein (2002) examines user perceptions of the quality of information found on the Internet using surveys of graduate and undergraduate students taking specific courses. The framework developed by Wang and Strong (Klein, 2001) was applied in the study as a tool for measuring data quality. The objective of the study is to further improve the understanding of users’ evaluations of internet IQ by comparing perceptions of graduate and undergraduate students. In a related study (Klein, 2001) compared user perception of information found on the Internet to data from traditional text sources. Klein’s focus, as ours, was on the consumers (users) of data and information. Klein’s study was built on prior research aimed at understanding the dimensions of data quality. The study has several limitations. First, the sample size was relatively small. Second, all of the surveyed users were students taking specific courses, particularly part of an MBA programme.

Third, respondents were asked questions about the quality of Internet and traditional text sources in general rather than being asked questions about specific Internet sites and text sources. Forth, our work examines how users perceived data stemming from multiple sources.

Our study addresses the quality of combined visual data. We aim to propose a set of principles of estimating data quality. The principles can subsequently be used to estimate these quality dimensions and present them to a user. Our studies research various flavours of data sources and focus on how people perceive subjective dimensions, specifically believability.

3 Experiment

3.1 Participants

The study was web-based, and was conducted through Amazon's Mechanical Turk. The study was open for about one week. 161 complete responses from 200 answers were identified. Some responses were excluded because participants had selected random numbers not within the two options provided, and all the answers where work time was less than 30 seconds were excluded. Participation was anonymous, and no information we stored could have been traced back to the participant. Each answer was paid $0.25.

3.2 Materials

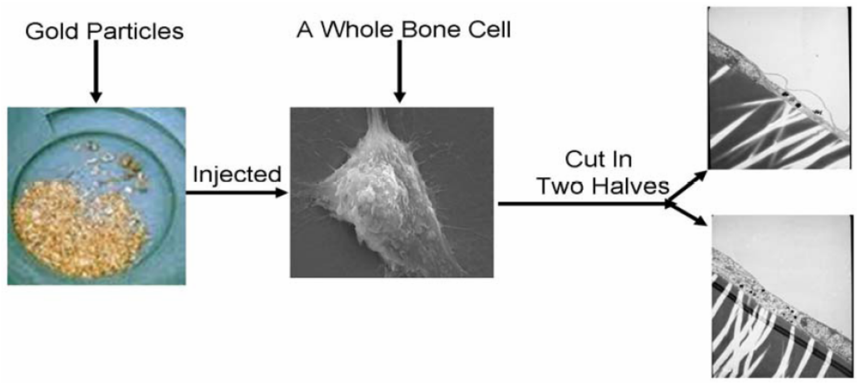
3.2.1 Data

Data was obtained from scientists in the Nanotechnology Centre at the University of Arkansas at Little Rock, and a sketch of the processed of incubation of the bone cells with gold nanoparticles is shown in Figure 1. The cells were sliced and visualised under a transmission electron microscope (TEM), where gold nanoparticles appear as black dots. The gold nanoparticles deposited on upper surface of cell plasma membrane, which triggers arms forming a round the gold nanoparticles (endocytosis). Some pictures used for the experiment show the arms in the process of endocytosis. Two different concentrations of gold nanoparticles are used 10 µg/ml and 160 µg/ml, but the end result of the incubation of the cells is the same regardless of the concentration.

3.2.2 Equipment and software

The software and environment used to perform the study is Amazon Mechanical Turk (Abuhalimeh et al., 2010), a marketplace in which people use their innate human intelligence to solve various tasks. The Mechanical Turk web service enables companies to programmatically access this marketplace, which is supported by a diverse, on-demand workforce. Mechanical Turk aims to make accessing human intelligence simple, scalable, and cost-effective. Businesses or developers that have tasks that cannot be solved by a machine, can create small pieces of work, called human intelligence tasks or 'HITs', via the Mechanical Turk APIs. Workers registered with the Mechanical Turk, then perform the tasks. Upon verifying the results, businesses and developers direct Mechanical Turk to pay the workers. We employed Mechanical Turk as a way to distribute questions about the gold-doped bone cells and to estimate the level of believability in the two gold concentrations from the professional workers registered with the Mechanical Turk.

Figure 1 Diagram describing the experimental process of incubation of the gold nanoparticles with the bone cells (see online version for colours)



Note: This image was provided to the participants in the study.

3.3 Methodology

The study was designed in such a way to not be dependent on accuracy. We achieved this goal by choosing a task based on the resulting cell configurations, which appears the same regardless of the gold concentration. The scientists discovered that the end-result of gold nanoparticles incubation is the same for both concentrations. However, scientists and experts were excluded from taking the study to avoid introducing bias towards accuracy in the results, since they would be familiar with materials and the images. This will not help the main goal of the study.

The first section of each HIT starts with short instructions about the HIT. The image shown in Figure 1 provides an overview of the whole process of adding the gold particles to the bone cells, and another image (Figure 2) provides a sample image with description of important features to allow the participants to familiarise themselves with the data types employed in the study. The second section of the HIT includes a textual description of the process of incubating the gold nanoparticles in the bone cells, and a sample of the two concentrations. The last section of the HIT describes the task the user needs to perform, and it is captured in Figure 3.

The study was broken down into nine different tasks (HITs). The first three HITs we designed included questions based on only images of bone cells doped with either 160 µg/ml and 10 µg/ml gold, while the next three HITs included text description of cells with each of the two concentrations. The final three HITs included both image and text integrated as in Figure 2. Different cells were presented in each HIT. All the HITs were published in a random order and at different times.

Figure 2 Snapshot showing the contents of images (see online version for colours)

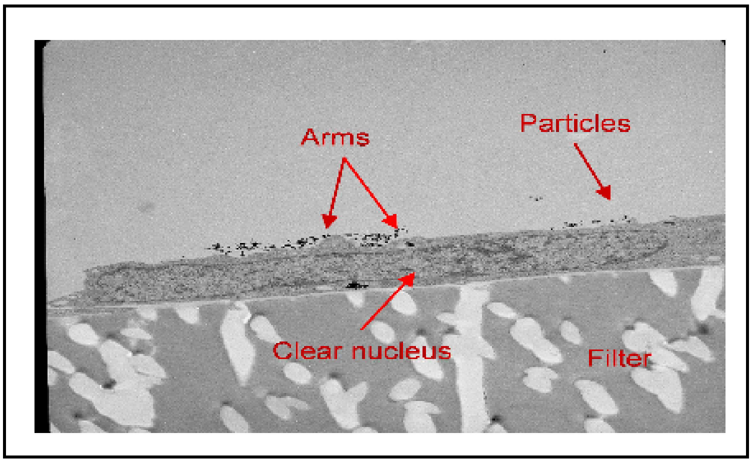
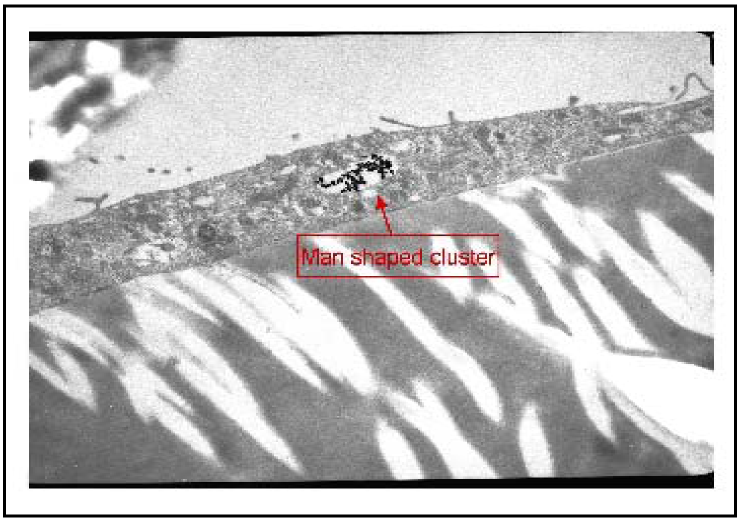


Figure 3 Snapshot showing a task that integrated images with text (see online version for colours)



Notes: Your job: look at Figure 3. John and Marta believe the concentration of particles applied to the pictures is 10 mg/ml, Mary and Jim believe the concentration is 160 mg/ml, which one do you believe?

In each HIT, the following scenario was included "John and Marta believe the concentration of particles applied to the pictures is 10 µg/ml, Mary and Jim believe the concentration is 160 µg/ml" as captured in Figure 3. Participants were asked to provide their answer whether they agree with John and Marta or Mary and Jim, and also to assess how confident they are in their answers on a five level rating scale. The scale presented the users with the following five choices: very confident (5), confident (4), neutral (3), not confident (2), and not confident at all (1). The time allotted per assignment was two minutes, and ten unique workers were allowed to work on each HIT. Only Mechanical Turk workers over 18 were allowed to work on the HITs. The payment for each assignment was $0.25.

3.4 Hypotheses

The following hypotheses were considered:

Hypotheses A	User's answers and believability does vary when showing image, text, or a combination.
Hypotheses B	Showing more pieces of information, combined information, improves the overall subjective assessment of believability.

3.5 Design

The independent variable in the experiment was source of data whose possible values are *image*, *text*, or *both*, and refers to the medium through which the participants in the study are getting their information. The textual information was extracted from the pictures in such a way to be similar to lab notes which present the features present in the observations (images). Note that actual concentration was another independent variable, but experts believe that it is not distinguishable in the images or text, and we do not consider the actual concentration as part of the model.

Two dependent variables were measured during the study believed_concentration and confidence. The believed concentration provides an objective assessment of the participant's believability and can be either 10 μg/ml or 160 μg/ml. The confidence is a self assessment from the user on a five level scale.

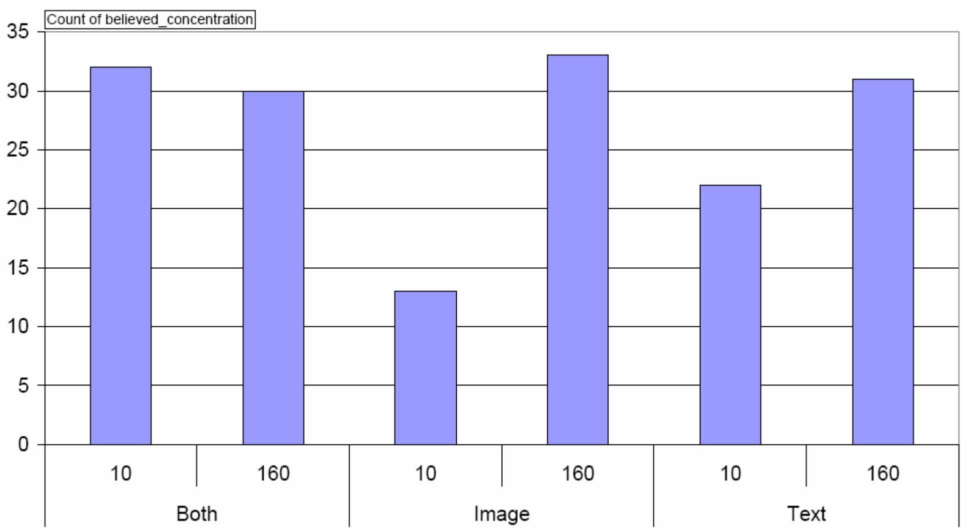
3.6 Results

The study was open for about one week, and 161 complete responses from 200 participants were identified. Some responses were excluded because participants had selected random numbers not within the two options provided, and all the answers where work time was less than 30 seconds were excluded.

An ANOVA revealed that source is a significant factor for believed_concentration (F2, 160 = 3.02, p = 0.0516).A Tukey pairwise comparison found significant differences between *image* and *both* (p = 0.0398). For confidence, the presentation medium is a marginal factor (F2, 159 = 2.64, p = 0.0748). Pairwise, *image* and *text* sources appear the most statistically different for confidence (p = 0.0597). Note that as expected, actual concentration is not a statistically significant factor.

Figures 4 to 6 illustrate the number of answers who believed either the 10 μg/ml or 160 μg/ml task. The self-assessment of the user confidence in their answers is given in Figure 7.

Figure 4 User believability in the two gold particle concentrations (see online version for colours)



Notes: The information is broken down by data source type and believed concentration. The y-axis shows the number of answers who believed in a given concentration.

Figure 5 Actual concentration of 160 µg/ml: user believability in the two concentrations by data type and believed concentration (see online version for colours)

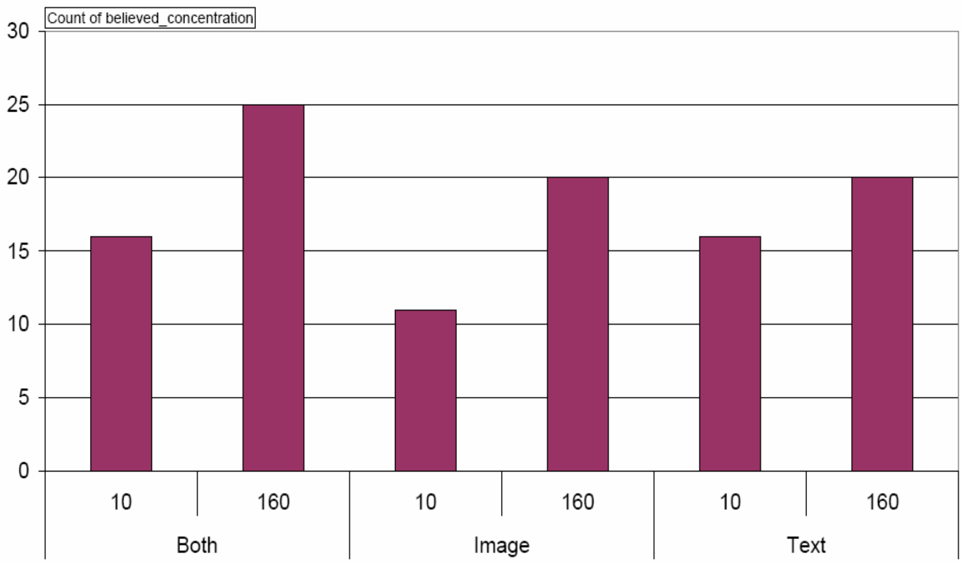


Figure 6 Actual concentration of 10 μg/ml: user believability in the two concentrations by data type and believed concentration (see online version for colours)

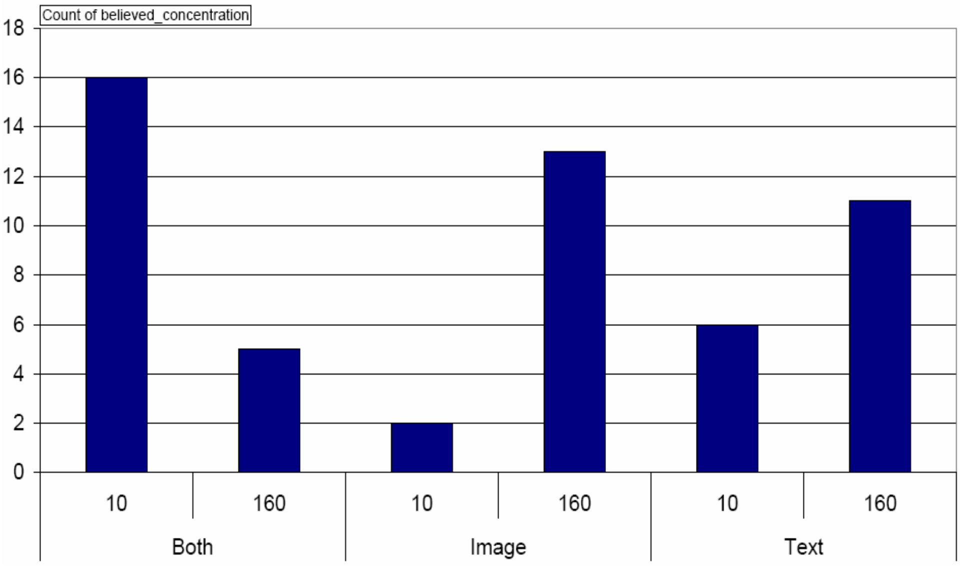
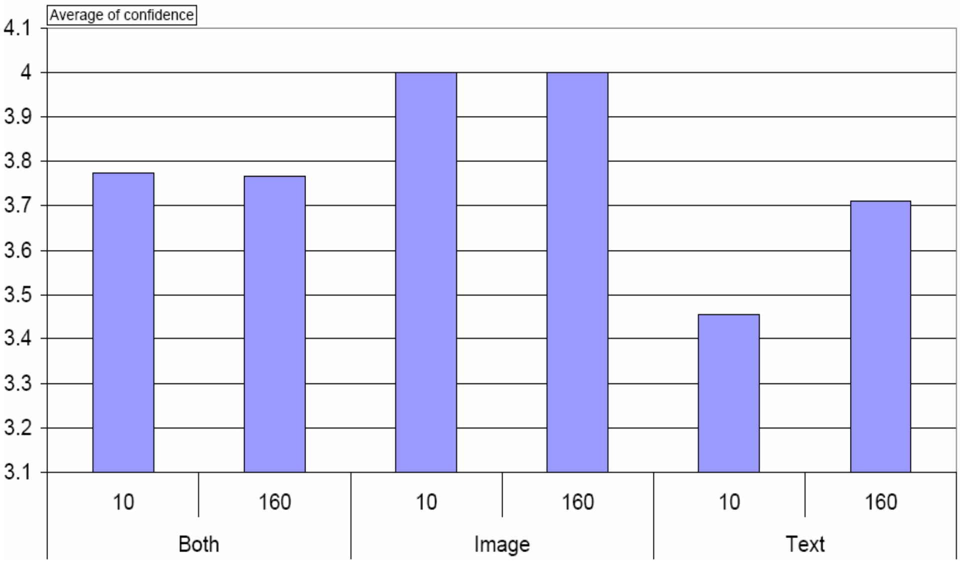


Figure 7 Average rating of users' confidence in their answers broken down by source type and believed concentration (see online version for colours)



Notes: 1 for user rating means 'not confident at all', 2 means 'somewhat not confident', 3 represents 'neutral', 4 represents 'confident', and 5 represents 'very confident'.

4 Discussion

Hypothesis A holds for both tasks for which the users were assessed, that is for both believing in 10 μg/ml and believing in 160 μg/ml. Hypothesis B holds only for the believability of 10 μg/ml task, as shown in Figure 4 more answers selected 10 for the *both* condition than for *image* or *text* alone. The believability of the 160 μg/ml is the lowest when users were presented *both* image and text combined, and thus Hypothesis B does not hold.

The results show a user preference (or bias) for the 160 μg/ml task, and consequently a bias against the 10 μg/ml. Figures 4 to 6 show that most people and under most conditions believed the concentration to be 160 μg/ml more than 10 μg/ml (except for the *both* case in Figure 6). Further research is needed to confirm the existence of this kind of biased.

Believability was task dependent in our experiment, which may make automated estimation of this dimension a complicated endeavour. A different behaviour of believability assessment is observed for the two tasks, none of them being simple averaging. The combination of the two datasets seems to affect slightly negatively the combination of text and images for the preferred task (160 μg/ml), while for the biased-against task (10 μg/ml) the combination improves the level of believability when compared to either *image* or *text*.

Most people are confident in their answers, which translates into them being confident in their belief. *Image* seems to inspire more confidence then *text*. For confidence, the combination of image and text produces a result that is about the average of the individual confidence levels as shown in Figure 7.

5 Relationship to other SIQ dimensions

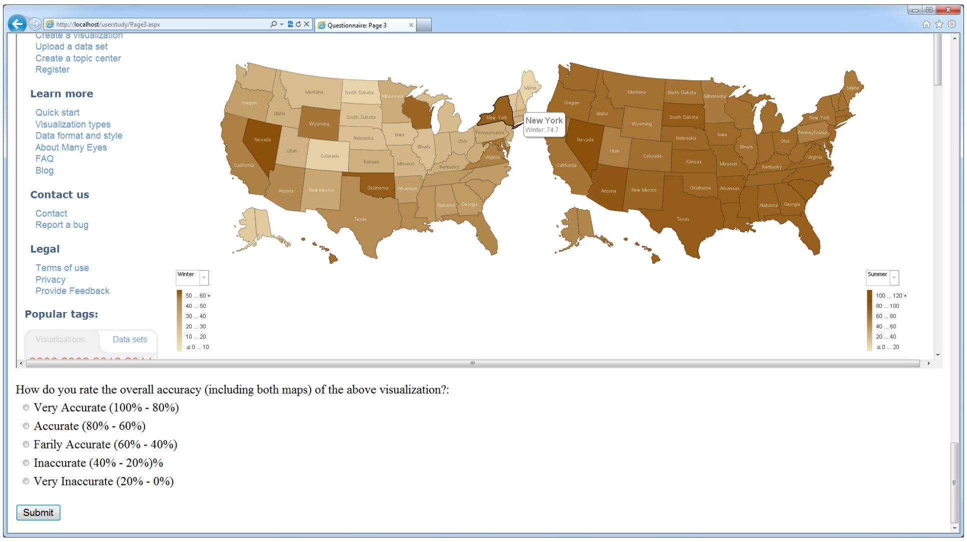
In a previous study (Abuhalimeh et al., 2010), user perception of accuracy was also found to be non-linear and task-dependent. The study was web-based, and was advertised to specific student groups in the information-related disciplines at the University of Arkansas at Little Rock and to colleagues of the authors. The study was open for about two weeks. Data was obtained from the National Oceanic and Atmospheric Administration (NOAA, 2010) and included average temperatures for all US states broken down by season. Only winter and summer were included in our study.

The accuracy of the data was artificially varied by selecting states at random and changing the reported average temperature. The new temperatures were randomly generated to fit within the minimum and maximum temperatures for that season that existed in the original dataset. As such, all ‘inaccurate’ values still fall within some reasonable limits. A special-purpose software was created for this task.

During the study, participants were presented with interactive visualisations such as the one in Figure 8, and were asked to assess the accuracy on a five level rating scale. Some visualisations were single panel, others were double. Participants were asked to rank the accuracy of the overall visualisation, which is not an issue for single panel views, but needed specific instructions for visualisation composed of two panels (that is two maps). The scale was presented as (1) very accurate (100%–80%), (2) accurate

(80%–60%), (3) fairly accurate (60%–40%), (4) inaccurate (40%–20%), and (5) very inaccurate (20%–0%).

Figure 8 Snapshot of a webpage showing a visualisation and question (see online version for colours)



How do you rate the overall accuracy (including both maps) of the above visualization?:

○ Very Accurate (100% - 80%)

○ Accurate (80% - 60%)

○ Farily Accurate (60% - 40%)

○ Inaccurate (40% - 20%)%

○ Very Inaccurate (20% - 0%)

Submit

Two independent variables were considered: *basic_accuracy*, and *additional_accuracy*. The basic accuracy is one of the 94%, 88%, 75%, or 50%, and represents the quality of the data presented in at least one of the panels of each visualisation. A webpage consists of four visualisation, one for each *basic_accuracy* value, shown in random order and affecting different US states.

Additional accuracy captures the quality of the data added in the case of double panel visual representations. The additional accuracy is a constant within each webpage, but it varies from webpage to webpage. The dependent variable are *user_estimation*, one of the five level scale answers as previously described, and *error*, which captures how far from the actual accuracy was the answer given by the user.

The following two hypotheses were considered:

Hypotheses A Average accuracy is a statistically important factor for user answer and error.

Hypotheses B1 Additional accuracy is a statistically important factor for user answer and error.

Hypotheses B2 more accurate additional data improves the overall subjective assessment of accuracy.

Hypothesis A and B1 are confirmed by the experiment, but B2, the addition of accurate data helps subjective assessment, is not supported by the results. One explanation for the failure to observe B2 is that more information contributes to a task overload and participants performed worse.

An ANOVA was performed for both the *user_estimation* and *error*. The *basic_accuracy* was found to be statistically significant factors: F10,110 = 2.01, p = 0.0385 for user estimation, and F10,110 = 2.26, p = 0.0192 for error. The same holds true for additional accuracy: F3,33 = 9.17, p = 0.0001 for user estimation, and F3,33 = 7.89, p = 0.0004 for error.

The results showed that visualisations with single maps were more accurately evaluated than double visualisations. Furthermore, there does not seem to be a linear dependence between the user assessment of accuracy and the actual, known accuracy. It may be possible that people have thresholds of how they perceive visualisations, and that they also suspect that an error has been introduced even for very accurate visualisations.

6 Future work

Assessing IQ is not an easy task and requires knowledge and awareness of the subjective and objective IQ metrics. Further studies may focus on additional tasks better understand the existence of preferred and biased against tasks. Such investigations may also need to be determined for other data types, and data presentation methods.

Subjective assessment is not limited to believability and accuracy, and our plans are to consider other SIQ dimensions and verify whether their behaviour is similar to the subjective accuracy and believability. Dimensions that are inherently subjective such as believability and value-added may lead to the development of a more complete theory of SIQ.

Any theory of SIQ may need to also consider the effect of data integration, an important topic in information and data quality. This study also showed that adding extra information is not always beneficial. Furthermore, for the cases when additional data is included, lower quality data may provide better support for subjective evaluation than higher quality data.

References

Abuhalimeh, A., Tudoreanu, M.E. and Peterson, E. (2010) 'Subjective evaluation of perception of accuracy in visualization of data', *Proceedings ICIQ Conference* (Little Rock, AR), pp.112–123.

Amazon, available at http://aws.amazon.com/mturk/ (accessed on 15 July 2010).

Bobrowski, M., Marre, M. and Yankelevich, D. (1999) 'A homogeneous framework to measure data quality', in *Proceedings of the International Conference on Information Quality (IQ)*, pp.115–124, Cambridge, MA.

Huerta, E. and Ryan, T. (2003) 'The credibility of online information', *AMCIS 2003 Proceedings*, available at http://aisel.aisnet.org/amcis2003/279.

Klein, B.D. (2001) 'User perception of data quality: internet and traditional text sources', *Journal of Computer Information Systems*, Vol. 41, No. 4, pp.9–15.

Klein, B.D. (2002) 'Internet data quality: perceptions of graduate and undergraduate business student', *Journal of Business and Management*, Vol. 8, No. 4, pp.425–432.

Lee, Y., Strong, D., Khan, B. and Wang, R. (2002) 'AIMQ: a methodology for information quality assessment', *Information Management*, December, Vol. 40, No. 2, pp.133–146.

Mo, L. and Zheng, H. (2008) 'A method for measuring data quality in data integration', *International Seminar on Future Information Technology and Management Engineering*, pp.525–527.

National Oceanic and Atmospheric Administration (NOAA) (2010) Available at http://www.nws.noaa.gov (accessed on 8 July 2010).

Nicolas, P. and Madnick, S.E. (2008) 'Measuring data believability: a provenance approach', Working papers 40086, Massachusetts Institute of Technology (MIT), Sloan School of Management.

Petty, R.E. and Cacioppo, J.T. (1986) *The Elaboration Likelihood Model of Persuasion*, Academic Press, New York.

Pipino, L.L., Lee, Y.W. and Wang, R.Y. (2002) 'Data quality assessment', *Communications of the ACM*, Vol. 45, No. 4, pp.211–218.

Wang, R.Y. and Strong, D.M. (1996) 'Beyond accuracy: what data quality means to data consumers', *Journal on Management of Information Systems*, Vol. 12, No. 4, pp.5–34.