Research About Measurability of Information Quality

Miloš Fidler and Dejan Lavbič(✉)

A.T. Kearney Svetovanje d.o.o., A.T. Kearney
(Global Management Consulting Firm), Dunajska cesta 151,
1000 Ljubljana, Slovenia
milosfidler@gmail.com, dejan.lavbic@fri.uni-lj.si

Abstract. This article will discuss ongoing research about Information Quality (IQ). Raters evaluating various IQ dimensions (accuracy, completeness, objectivity…) of same object showed low agreement level, therefore making IQ not measurable. Increase of IQ measurability to sufficient level would present an opportunity for guidelines to replace information of low with high quality. Speculations why IQ dimensions are not measurable have been made but at the same time mechanisms that improve agreement level have been proposed by researchers for validation. Moreover context in which information is being evaluated has not been yet addressed by existing research. This article will describe and explain a study that aims to create a robust model that will validate and measure effect of three different IQ aspects. Although this article is still work in progress, current results regarding research construction and preliminary testing will be presented as well as future steps.

Keywords: Information quality · Inter rater reliability · Information quality dimensions · Completeness · Accuracy · Objectivity · Consistency · Big data

1 Introduction

Information quality (IQ) remains one of the most intriguing fields of computer information science research [1, 2]. While economies, companies and our daily activities are becoming more and more data driven requirement for data and information quality is increasing. Digitalization and ongoing progress in technology established possibilities to generate and store huge quantities of data [3]. But big data by itself does not necessary also mean high quality [4]. Therefore questions: what is IQ and how are information of high quality produced became a challenge.

Arazay and Kopak in [5] state that research focusing on IQ concept, has been investigated extensively [6, 7] to identify underlying dimensions (or attributes) of IQ, such as accuracy, completeness, presentation, and objectivity which are important to information consumer. Studies [8, 9] outlined importance of perception about IQ – information consumers may perceive certain quality dimensions to be more important than are others, and for a variety of reasons, including domain expertise. Moreover Arazay and Kopak [5] proved that IQ consumers perceive IQ so differently that when being asked to rate same information (Wikipedia article), rates between consumers

L. Uden et al. (Eds.): KMO 2015, LNBIP 224, pp. 272–281, 2015.
DOI: 10.1007/978-3-319-21009-4_21

varied so much that it was not possible to determine if information is of high or low quality. Studies [1, 10] have managed to define the basic IQ components (answering partially what are IQ components) but did not yet answer how information of high quality should be produced.

Research focus was shifted to following two questions: (1) how would IQ be made more measurable (2) and why is IQ not measurable?

If IQ is not measurable then no rule can be made to produce information of higher and lower quality. However if IQ would become measurable then information of higher quality would provide a potential to determine a rule or a standard. Understanding why IQ is not measurable could on the other hand identify key activities to improve measurability and avoid critical situations which cause measurability to be limited.

2 Measurability of IQ

Measurability challenge is common to rating systems, for example rating student's performance in University. Oakleaf [11] argued that rating system becomes objective when all raters give to same student identical or at least similar mark. Objectivity can be further increased by pre-learning coordination groups where raters rate different examples and discuss why a certain mark is most appropriate. Additionally, Okleaf calculated that agreement level varies between rating groups (librarians had overall lower rating agreement level as English teachers).

Jens-Erik Mai is in his latest research [12] confirming this argument: IQ is context-dependent, and can only be assessed and understood from within specific situation and circumstances and therefore remains centered on the meaning of information and the understanding of it in a society. Wang and Strong [7] described IQ with term fitness for use - how much consumer benefits from use of information in a specific situation. Hjorlad [12] was in his study evaluating four articles about breast cancer. He was wondering about his preferences since he was the one also rating articles. What if it would not be him assessing the articles but for instance a grade five student working on a school project? Would grades still be the same when the needs of both raters are completely different?

Eppler [13] pointed out that apart the degree to which the information at hand either meets the requirements of the particular activity in which the user is engaged (the objective view) there is also an subjective view, the degree to which the information meets the expectations of the user.

It could be that existing research measured low rater's agreement level due to "loose" definition of context in which information is rated. We are arguing that most of the research [5] was asking interviewees to rate IQ purely on hypothetical basis, giving users absolutely free hands to interpret the "why am I rating" question. This could cause a situation in which two very similar raters with otherwise high agreement level produced additionally dispersed rates due to different interpretation of context. Let's assume that two mathematical teachers with similar working experience would be asked to rate same math exams. The first would be told that his rates are purely hypothetical, he should just rate it as he feels right but second needs to rate exams right since his rates will be reviewed by national rating commission and linked to incentive scheme.

Reijers and Mendling [14] pointed out another possible reason for disagreement between raters: personal and subjective factors. In principal when adjusted to IQ topic these factors mean how a rater perceives or understands the instruction of rating. Jens-Erik Mai [12] was referencing to them as social or cultural effects, arguing that user with different interest will differently understand rating question, information and rating context. We argue that rating is a decision making process that can't be measured directly but mostly interpreted indirectly by rater characteristics.

Reijers and Mendling [14] found out that "persona" factors had higher explanatory power then model related factors. In perspective of IQ this would mean that rater's rate can be better predicted by his/hers characteristics then by the information and research context.

To summarize, to date little is known about the interrater reliability of IQ dimensions, and we can only speculate which factors effect it most. Our investigation aims to fill this gap, and our research questions which aspects effect most the rater agreement level. Moreover we hope also to calculate the overall effect of personal characteristics additional validating Reijers and Mendling [14] research at area of IQ.

3 Research Methodology

Our research focuses on underlying questions: why is IQ not measurable and how much more measurable can it become?

In order to achieve our goal we have constructed a robust model that will help as analyze three different aspects effecting measurability of IQ: rating inputs and mechanics (questioner's basic inputs and mechanics), rater's characteristics (that have potential to correlate with decision making of choosing a rate) and situation (context in which rater is evaluation information).

Each aspect will be addressed with unique approach and further clarified.

3.1 Rating Inputs and Mechanics

We are arguing that agreement level of IQ raters can be further increased as of teachers in Oakleaf [11] in a way how questions and objects in study are constructed. Therefore we have carefully analyzed rating process and defined input factors affecting ratings and mechanisms with a potential to increase agreement level.

We have defined two groups of factors and mechanisms: factors and mechanisms that were already mentioned in research but require validation (i) and purely hypothetical, not yet addressed factors and mechanisms (ii).

First group will address rater's familiarity with the topic of which objects are being rated. We will also investigate how ratings are affected with number of rated instances (learning effect) and calibration, an option to compare already given rates by objects and change them (revision effect).

Second group will address a form in which information is presented to consumer. Most of the current research [2, 5] was performed on Wikipedia articles of various size and topics. We will consider cases of rating where objects are prepared in txt only

format, picture only or combination of both. Additionally, complexity of objects being rated will be varied from article summaries to simple statements. Lastly, raters rating behavior will be tracked. We will measure how much time it takes for rater to perform ratings and how many mark corrections are made.

We hope to validate rating inputs and mechanics from first group and determine the size of its effect on measurability. We expect that second group of rating inputs and mechanics will also have an effect on measurability and that effect's size can be calculated as well. Moreover we believe that both effects combined will have a significant effect on IQ measurability.

3.2 Rater's Characteristics

We will also address aspect of rater's characteristics. Rating by its nature is a decision making process where in a simplified manner different people decide differently. We hope to clarify both differences.

First, key rater characteristics with highest effect on ratings will be identified. Second, segments and groups of raters with lowest and highest agreement level will also be identified. Third, we will see which rater's characteristics have a potential to maximize agreement level within segments.

To achieve this we will try to obtain different questioner feedbacks from people of various professions (not only MBA students, teachers and librarians [5]) by leveraging our networks. Various respondents will enrich our research and will make it more applicable to real life.

Moreover we have constructed several questions that will help us identify differences between raters and open options for segmentation. Characteristics have been organized in four groups: geographic (location), demographic (age, gender, family size…), psychographic (personality, hobbies…) and behavioral (rater's relationship towards IQ).

3.3 Situation or Context

Additionally we will also upgrade existing research by setting context for raters.

Apart from existing research [2, 5] where respondents are asked about IQ purely on research basis with no further clarification of the situation in which information evaluated is used, we will also simulate two other cases.

First case will simulate how raters will behave in situation where their rating skills are being checked (to obtained golden rule). User will not only be provided with the object that is being rated but also with the golden rule standard. We hope to see how well will the rater identify and evaluate the difference from the golden rule. We believe that golden rule situations will make raters more responsible since they will have feeling that their knowledge is being checked, resulting in higher pressure and extended rating times. Agreement level should therefore increase.

Second case will simulate how raters will behave in a situation where evaluated object is used for the task at hand. We believe that tasks will make raters even more

focused and their judgment more objective causing agreement level to increase. Moreover we believe that there will be a significant correlation between rater's evaluation of information quality and his success at solving task. We presume that rater's performance at tasks will significantly impact his rates.

3.4 Research Construction

A set of experiments was developed to test, validate and establish the interpreter reliability of above described aspects. Basic research settings have been unified with existing studies [5] to make results comparable as much as possible. Each research study has therefore been based on same set of assumptions (number of rating scale values, rated IQ dimension…) that have already been used in existing research.

Experiments were organized in three research studies each addressing unique situation or context (see Fig. 1).

Research I is simulating same "academic" context settings as in existing research where rater was not provided with any context apart from purely hypothetical - what he feels is the right answer. We will use this to establishing a baseline comparable to existing research (cross validation) and by also determining the significance of aspects on agreement level.

Research II is imperceptible putting rater in a situation where his knowledge is being checked. It was designed as upgraded version of research I, where effects of aspects are measured and complexity level of information to-be rated reduced. Moreover it kind of replicates real life rating situations where raters are challenged to rate randomly presented cases. We hope to validate accuracy of these systems by measuring the effect of learning on agreement level trying to found out how many examples must a rater rate to achieve certain stability. Another important validation

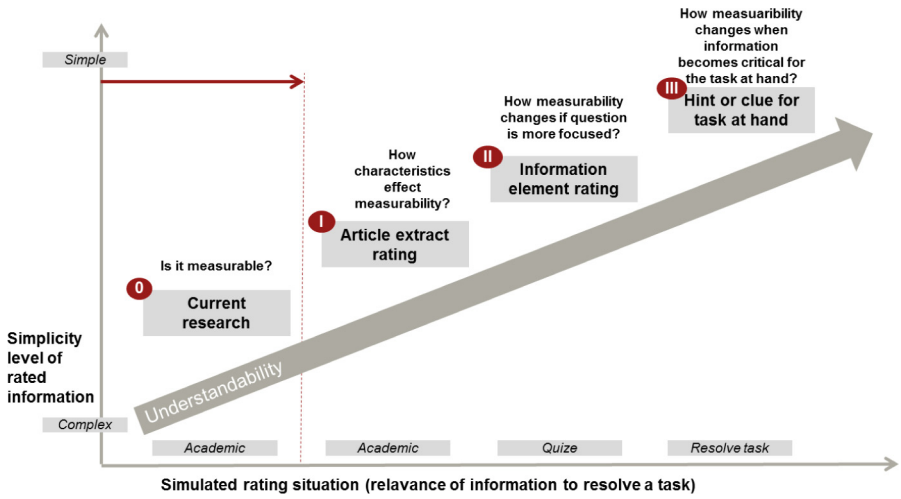

Fig. 1. Study design (self-made sketch)

mechanism will be the calibration. We will measure how selected cases and their order effect rater's requirement for additional calibration of already given rates.

Research III is further upgraded research II even more simulating real-real live examples and complexity level of information simplified. User can only rate true value of information when rated object is critical for the task at hand. We have therefore designed a special game with set of tasks. Each task consists of riddle where information is provided to a user as a hint for solution. User can then on basis of IQ and his thinking cognitive capabilities less or more successfully find a solution. Successful users are rewarded by more points, best results published as high scores and specially made storyline is motivating user not to quite. User rates the IQ of the hint before he/she tries to solve the puzzle and there is another option to change the rating after the riddle is resolved.

3.5 Research Implementation

Our study was implemented by agile approach: incremental cycles of study (product) development and testing (to collect rater requirements).

Advance research settings such as agreement mechanisms, assessed information and research mechanics were in first stage designed on common sense. We have brainstormed several possibilities and based on common sense developed study specification considering ourselves as alfa-testers.

Study was then implemented in java script as a web page questioner where beta-testing was performed by carefully selected individuals not being yet tainted with IQ research. Our crucial criteria here was the execution feasibility – is the questioner self-explanatory and short enough so that interviewee won't stop half way through. Feedbacks have been collected and implemented study further upgraded made more user friendly, easier (less clicks, more intuitive) and more fun (advance visuals) to fill.

We hope to obtain at least 300 completely solved researches of each type from Slovenian and additional 300 worldwide. Our initial target group of interviewers will be students of Faculty of Computer and Information Science. We also hope to get significant response from secondary schools. Additional interviewers will be obtained either by associations, coworkers, friends.

Apart from current research we hope to have our raters different as possible since we believe that this additionally effects agreement between them. Therefore we plan to translate the study in English and use Amazon Mechanical Turk for rater recruitment. We will also test the attractiveness level of research level III by offering it to raters as a stand-alone questioner.

4 Preliminary Results

In this chapter we will present our current survey results. First part of the chapter will focus on overall surveys performance (research I, II, III) and second on performance of rated articles in research I. Consider that this is still work in progress we will focus mostly on results of research I since we believe it is most stable. There is great change

that numbers will change, since we are currently working with small sample of similar interviewees (same background).

Until this moment we have obtained 79 completed surveys for research I & II and additional 26 for research III. The difference in completed surveys occurred due to parallel roll out of research I & II, while implementation of Research III was still in progress.

Our interviewees have been students of 2nd year study of undergraduate university program performed on Faculty of Computer and Information Science, University of Ljubljana.

4.1 Survey Performance

We can say that it paid off to put extensive amount of time and effort in research design and testing due to several reasons.

First, we achieved high completion rates with minimal drops. Only 2 interviewees out of 81 have not complete Research II (98 % of completion rate). We have received 100 % completion rates for Research I & III with no drops;

Second, short survey completion times were achieved (Fig. 2). Interviewees have completed all three researches in app. half an hour (31:14 s). The largest amount of time has taken the completion of Research III with on average 19:09 s and more than 60 % of total time spent (for the interviews answering all research questions).

Third, we also managed to achieve efficient use of interviewee's time. Interviewees spent on average app. 39 s per question, making a survey very interactive (Fig. 3). Completion times where as expected progressively lengthened since each research has intentionally been designed as an upgrades of previous one to leverage interviewees learning about survey mechanics to full extent. Apart from more question, upgrades also improved context of research to be more and more aligned with real-life situations. Therefore time per question did not increase from research I to research II, since

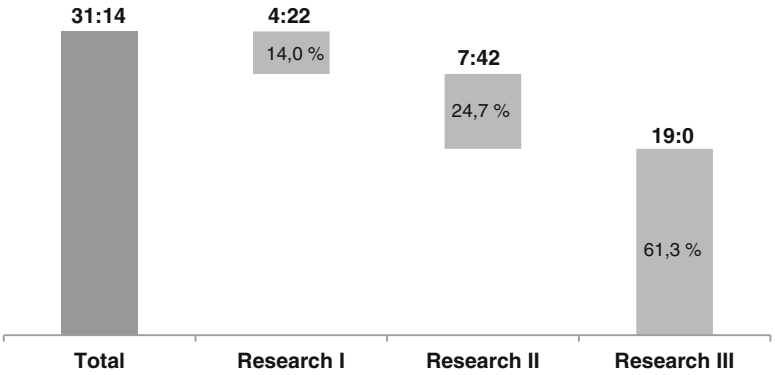

Fig. 2. Total time spent per research; Interviewees have spent on average 31:14 s to finish complete package of all three researches. The longest time to complete took research III with 19:09 s. Research II took with 4:22 s app. twice as long as research I with 7:42 s.

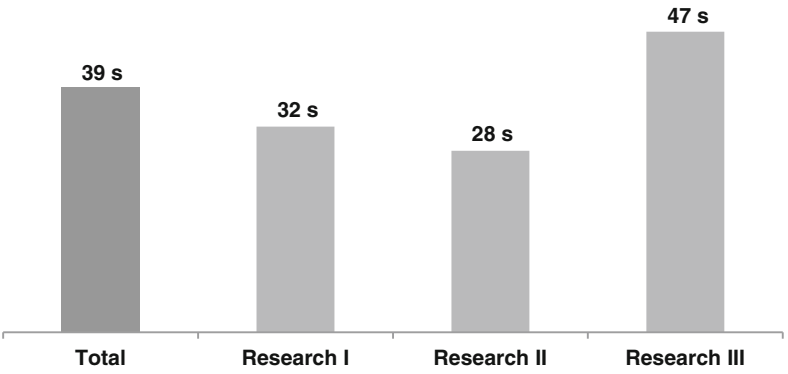

Fig. 3. Average time spent per question; Interviewees have spent on average app. same amount of time to answer a question in research I & II, app. 30 s. Time spent per question in research III exceeded time spend per question in research I & II for 60 %, resulting in 47 s. Most of the difference can be explained by interviewees solving riddles.

simulated situation did not burden interviewee with more work. However this has not been the case for research III where amount of time spend per question increased for 17 s (60 %) compared to research I & II. Interviewees have spent more time per question due to resolving riddles (simulation of task-at-hand situation).

4.2 Research Implementation

We have used in research I following articles: homo-sapiens (human), Occitan (archaic french language), alexandrite (valuable gem stone). Articles have been made as summaries of various online sources (Wikipedia, forums, encyclopedias) with same structure and length of around 180 words and 14 sentences. Each article has been structures into parts like definition, discovery, characteristic… so that topic was presented what we believed to be the most concise and efficient way. Therefore we had every reason to believe that articles will receive good marks.

Interviewees rated articles as good, they have in general agreed that overall quality of articles is slightly better than neutral (Table 1). Interviewees rated with best overall IQ mark article about alexandrite (1,28) and with the lowest overall IQ mark article about homo-sapiens (1,00). Article about alexandrite exceeded average marks in all IQ dimensions, apart from accuracy. Alexandrite was also the only article that interviewees found complete.

Overall from IQ dimension prospective interviewees rated best objectivity, giving it a very good mark. Consistency of representation received good mark, while completeness and accurate received a mark between neutral and good.

Table 1. Average grades of rater articles per IQ dimension; Raters were grading articles with 7 unit Likert, having 3 not agree, one neutral and 3 agree units. When calculated averages we mapped not agree values from -3 to -1, natural as 0 and agree values from 1 to 3.

IQ dimension	homo sapiens (human)	occitan (language)	alexandrite (gem stone)	Total
accuracy	0,85	0,28	0,70	**0,61**
completeness	0,22	0,89	1,26	**0,79**
objectivity	2,19	1,96	1,79	**1,98**
consistency of representation	0,74	1,21	1,36	**1,10**
Total	**1,00**	**1,08**	**1,28**	**1,12**

5 Conclusion and Future Works

We are implementing a three part study about IQ that will help us build a robust model and determine effects on IQ agreement level of three different aspects: rating inputs and mechanics, rater's characteristics and situation. Our study was designed in a way to add as much as possible of aspect elements neglected by current research but also to have at the same time sustainable response times. Additionally, each research faces rater with different situation in terms of usefulness of evaluated information.

Hopefully robustness of the model will help us determine how agreement level is affected by each aspect and what the scientific and practical implications are. We hope that our model will show how effects may vary in different situation for various rater segments. Moreover, we are very interested how overall agreement level (and of certain segments) will evolve throughout different researches.

At this stage we are still testing and improving our questioner. It seems that we managed to design user friendly survey with a great potential for academic contributions. We have achieved preliminary validation of rated article summaries in research I and saw that they have on average obtained slightly better mark than neutral. We hope to obtain more responses so that we are able to build IQ agreement level model and compare results with Arazay and Kopak's research [5].

References

1. Hilligoss, B., Rieh, S.Y.: Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. Inf. Process. Manage. **44**(4), 1467–1484 (2008)
2. Yaari, E., Baruchson-Arbib, S., Bar-Ilan, J.: Information quality assessment of community-generated content - A user study of Wikipedia. J. Inf. Sci. **37**(5), 487–498 (2011)
3. Ge, M., Helfert, M.: A review of information quality assessment. In: IET Conference Proceedings, Institution of Engineering and Technology, pp. 951–958 (2007)
4. Geiger, J.G.: Data quality management: the most critical initiative you can implement. In: SUGI 29, Montreal, Canada (2004)

5. Arazy, O., Kopak, R.: On the measurability of information quality. J. Am. Soc. Inform. Sci. Technol. **62**(1), 89–99 (2011)
6. Rieh, S.Y., Danielson, D.R.: Credibility: a multidisciplinary framework. Annu. Rev. Inf. Sci. Technol. **41**, 307–364 (2007)
7. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. **12**(4), 5–33 (1996)
8. Flanagin, A.J., Metzger, M.J.: The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. New Media Soc. **9**(2), 319–342 (1981)
9. Lee, Y.W., et al.: AIMQ: a methodology for information quality assessment. Inf. Manag. **40**(2), 133–146 (2002)
10. Liu, Z.M.: Perceptions of credibility of scholarly information on the web. Inf. Process. Manage. **40**(6), 1027–1038 (2004)
11. Oakleaf, M.: Using rubrics to assess information literacy: an examination of methodology and interrater reliability. J. Am. Soc. Inform. Sci. Technol. **60**(5), 969–983 (2009)
12. Mai, J.E.: The quality and qualities of information. J. Am. Soc. Inform. Sci. Technol. **64**(4), 675–688 (2013)
13. Eppler, M.J.: Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes, 2nd edn. Springer, Berlin (2006)
14. Reijers, H.A., Mendling, J.: A study into the factors that influence the understandability of business process models. IEEE Transact. Syst. Man Cybern. **41**(3), 449–462 (2011)