Journal of the
Association for Information Systems

Research Paper ISSN: 1536-9323

Heuristic Principles and Differential Judgments in the Assessment of Information Quality

Ofer Arazy
University of Haifa, Israel
oarazy@is.haifa.ac.il

Rick Kopak
University of British Columbia, Canada

Irit Hadar
University of Haifa, Israel

Abstract:

Information quality (IQ) is a multidimensional construct and includes dimensions such as accuracy, completeness, objectivity, and representation that are difficult to measure. Recently, research has shown that independent assessors who rated IQ yielded high inter-rater agreement for some information quality dimensions as opposed to others. In this paper, we explore the reasons that underlie the differences in the "measurability" of IQ. Employing Gigerenzer's "building blocks" framework, we conjecture that the feasibility of using a set of heuristic principles consistently when assessing different dimensions of IQ is a key factor driving inter-rater agreement in IQ judgments. We report on two studies. In the first study, we qualitatively explored the manner in which participants applied the heuristic principles of search rules, stopping rules, and decision rules in assessing the IQ dimensions of accuracy, completeness, objectivity, and representation. In the second study, we investigated the extent to which participants could reach an agreement in rating the quality of Wikipedia articles along these dimensions. Our findings show an alignment between the consistent application of heuristic principles and inter-rater agreement levels found on particular dimensions of IQ judgments. Specifically, on the dimensions of completeness and representation, assessors applied the heuristic principles consistently and tended to agree in their ratings, whereas, on the dimensions of accuracy and objectivity, they not apply the heuristic principles in a uniform manner and inter-rater agreement was relatively low. We discuss our findings implications for research and practice.

Keywords: Information Quality, Assessment, Heuristic Principles, Consistency, Rating, Agreement.

Jeffrey Parsons was the accepting senior editor. This paper was submitted December 21, 2013, and went through 3 revisions.

1 Introduction

Research on readers' ability to recognize quality information as they encounter it has grown in importance over the last decade or so. Largely, the increasing importance results from several factors, including the current dominance of the Web as a primary source for information in all its forms, the heterogeneous nature of this information, and the Web's almost instantaneous and universal accessibility (Yaari, Baruchson-Arbib, & Bar-IIan, 2011). One result of this rapid expansion in the amount of information available is a parallel diminution in the proportion of information that benefits from traditional gatekeeping processes on the "information-production" side (Metzger, 2007). As a result, some researchers have expressed concern for the quality of content on the Web, particularly in areas such as health information (Eysenbach, Powell, Kuss, & Sa, 2002; Gagliardi & Jadad, 2002) and increasingly with user-generated content (Lukyanenko, Parsons, & Wiersma, 2014). A growing proportion of information available is not—and realistically cannot be—subject to peer-review or some other rigorous and transparent vetting process. This situation has led to a condition that Lankes (2008) calls "information self-sufficiency": that is, the condition in which consumers themselves face more responsibility to judge information's quality. This disintermediation has resulted in somewhat of a conundrum: as information consumers take on more responsibility for assessing the quality of information encountered, they increasingly need to do so based solely on broad, structural characteristics of the information itself (and on the presence of cues that represent these characteristics) (Lankes, 2008).

With this need for greater self-sufficiency, the factors that users employ in making judgments about the quality of the information they encounter become an important focus of research. Much of this research focuses on exploring the underlying dimensions of information quality (IQ), such as accuracy (factual correctness), completeness (inclusion of all relevant information), objectivity (the lack of bias), and representation (clear, concise, and consistent presentation) (Eysenbach et al., 2002; Hilligoss & Rieh, 2008; Kim, Eng, Deering, & Maxfield, 1999; Lee, Strong, Kahn, & Wang, 2002; Liu, 2004; Rieh & Danielson, 2007; Wang & Strong, 1996). To a large extent, these studies have focused on the efficacy of the various dimensions to determine whether one particular IQ dimension better represents users' perceptions of information quality than the others. These studies reveal that information users may view some quality dimensions to be more important than others, such that contextual factors such as domain expertise (Stanford, Tauber, Fogg, & Marable, 2002), gender (Flanagin & Metzger, 2003), or differences in information-seeking style (Rains & Karmikel, 2009) may influence users' perceptions.

Pivotal to our study is the question: what causes users to vary in how they assess IQ? Recently, research has proposed that the differences between IQ dimensions in terms of inter-rater agreement may be associated with the cognitive processes that occur when individuals assess IQ. For instance, Arazy and Kopak (2011, p. 92) propose that "differences in inter-rater reliability between various dimensions may stem from the availability of cues or the application of heuristics". In other words, given a particular context, some quality dimensions may be more and others may be less amenable to the application of heuristic principles. Gigerenzer and Todd (1999) refer to heuristic principles as the "building blocks" with which individuals construct specific heuristics. They propose three heuristic principles that govern the means by which individuals "search" for relevant cues in the information space, "stop" the search for additional cues, and "make decisions" based on the cues found. As they state, "These heuristic principles are the building blocks, or the ABCs, of fast and frugal heuristics" (Gigerenzer & Todd, 1999). Focusing on the process by which independent information consumers assess the quality of resources, we examine the association between the consistent application of heuristic principles and inter-rater agreement (or disagreement) in the rating of the quality of user-generated content. The setting for our empirical investigation is Wikipedia, an exemplar of peer-production (Benkler, 2006) and one of the most popular websites today (http://www.alexa.com/topsites). The contention regarding the quality of Wikipedia articles (Giles, 2005) and the fact that previous research on information quality has used it as its setting (Chesney, 2006; Fallis, 2008; Lim, 2009; Luyt, Aaron, Thian, & Hong, 2008; Stvilia, Twidale, Smith, & Gasser, 2008) makes Wikipedia an ideal setting for our investigation.

To consider the potential association between the consistent application of heuristic principles and inter-rater agreement, we conducted two studies. In the first study, we qualitatively investigated the cognitive process by which participants—university students and librarians—assessed the quality of Wikipedia articles along four dimensions of IQ: accuracy, completeness, objectivity, and representation. We paid particular attention to participants' application of a set of heuristic principles and examined the consistency with which they applied these principles when assessing each of the IQ dimensions. In the second study, we recruited three university librarians as participants to assess the quality of a larger set of Wikipedia articles and measured inter-rater agreement in IQ assessment along the dimensions of accuracy, completeness, objectivity, and representation and in the number of errors and omissions in each article they identified. With the results from these two studies, we identified a possible association between divergence in the application of heuristic principles and inter-rater disagreement on IQ assessments.

Understanding the role of heuristic principles in determining which dimensions yield more or less agreement in IQ assessment has significant value for both research and practice. Prior research in the information systems and information science fields has paid little attention to the cognitive processes underlying the assessment of information quality. Our findings add validation to the building blocks framework (Gigerenzer & Todd, 1999) and demonstrate the usefulness of this conceptualization to research on IQ assessment. The implications for practice include recommendations for information consumers who rate online content and for Web services that produce information quality metrics for published content.

This paper proceeds as follows: in Section 2, we first review background literature on the dimensions of information quality and on the issues that surround the assessment of these dimensions. In Section 3, we review the theoretical context and describe the role of heuristics in decision making. Using Gigerenzer’s building blocks framework (Gigerenzer & Todd, 1999; Gigerenzer et al. 2011), we develop our argument regarding the relationship between the consistency in the application of heuristic principles and the resulting inter-rater agreement (or lack thereof) in the assessments of IQ. In Section 4, we describe our method for investigating how the consistent application of heuristic principles determines the measurability of IQ dimensions and report the results of our empirical studies. In Section 5, we present our results. In Section 6, we elaborate on our findings’ implications for research and practice, note some limitations of our study, and provide some possible avenues for future research. Finally, in Section 7, we conclude the paper.

2 Assessing the Quality of Information

Information quality is hard to define (Michnik & Lo, 2009); it is “elusive…[and] of a transcendent quality (essence) synonymous with excellence” (Fink-Shamit & Bar-Ilan, 2008). Hilligoss and Rieh (2008) stress the importance of users’ information and view information quality as the individuals’ “subjective judgment of goodness and usefulness of information” (p. 1469). Alternatively, Taylor and Voigt (1986) see the quality of information as its value in relation to the purposes for which one uses it. From a more utilitarian perspective, we might also acknowledge the “objective” and “subjective” views of information quality as Wang and Strong (1996) do in their definition of data quality[1]. For example, Wang and Strong (1996) use the phrase “fitness for use” to represent the importance of context and the manner in which one's assessment of quality depends on the “fitness” of the data to one's specific assessment purposes.

As one might expect, there are also variations in the nomenclature used to operationalize such a multi-dimensional construct. Typically, similar attributes of information quality are sorted into higher-level groups, or “quality dimensions”, and ascribed a representative name. The sorting mechanisms vary from the application of intuitive, pre-determined, top-down classification schemes, to reliance on formal, statistical procedures such as factor analysis. Taylor and Voigt (1986), for example, identified five kinds of value (i.e., dimensions) that comprise information quality: accuracy, comprehensiveness, currency, reliability, and validity. Alternatively, Wang and Strong (1996) identified data quality dimensions through studies of user-based descriptions of quality. Several reviews have attempted to create information quality typologies based on these empirical studies. Lee et al. (2002), for example, collected IQ attributes from fifteen prior studies and, adapting the categories that Wang and Strong (1996) propose, reduced the information quality attributes to four main categories.

Although such studies succeed in reducing the number of information quality dimensions to more manageable numbers, their variety remains substantial. In the investigation at hand, we focus on a reduced set of these dimensions rather than attempting to cover their full range. As we state above, we explore the association between the consistent application of certain cognitive heuristic principles used in assessing the various IQ dimensions and the inter-rater agreement levels between these assessments. To make the study more manageable in this regard, we used the same quality dimensions that Arazy and Kopak (2011) employed to study information quality measurability: accuracy, completeness, objectivity, and representation. These dimensions have been used in other studies of IQ (West & Williamson, 2009) and in meta-analyses of

[1] We do note in reference to Wang and Strong (1996) that they are speaking of data quality as distinct from information quality. However, for our investigation of quality dimensions and their assessment, data quality and information quality share sufficient similarities (Knight & Burn, 2005; Nurse, Creese, Goldsmith, & Lamberts, 2011); hence, we consider the findings from Wang and Strong (1996) to be relevant to our purpose.

the health informatics field (Eysenbach et al., 2002; Kim et al., 1999). We do not argue that these quality dimensions are necessarily more important than others; rather, we argue that this subset reasonably represents the different kinds of information quality dimensions that others have viewed as important and that researchers have employed these same dimensions with success when studying similar issues.

There are growing concerns regarding users' ability to recognize quality information when they see it. For example, Wikipedia has recently begun rating the quality of its articles with a set of predefined quality categories. Yet, the lengthy discussions in Wikipedia on the procedure for determining articles' quality (Stvilia et al., 2008) demonstrate how difficult it is to come up with consistent and objective quality assessment criteria. Indeed, the research community has increasingly begun to pay more attention to the measurability of IQ (Yaari et al., 2011). An important aspect in discerning information quality is the extent to which independent assessors agree on the quality of a particular information element. In recent years, studies in various fields have considered inter-rater reliability when assessing the quality of information. For example, Moskal (2000) discusses scoring rubrics that educators use for evaluating students' work in primary, secondary, and college-level education and considers inter-rater reliability; LeBreton and Senter (2008) review the issues surrounding the use of inter-rater reliability in organizational research; and Oakleaf (2009) discusses the rubric-based approach to assessing information literacy and stresses the importance of inter-rater reliability.

Different information quality dimensions present varying challenges in terms of assessment. However, we still know little about the degree to which multiple assessors tend to agree about the quality of information when asked to judge the same information. Arazy and Kopak (2011) shed some light on this issue by comparing agreement levels among university students who analyzed a relatively large set (close to 100) of Wikipedia articles. They found that overall inter-rater reliability levels were lo, and that the dimensions of completeness and representation yielded higher agreement levels than did the dimensions of accuracy and objectivity. They then speculated that differences in the heuristics employed may have accounted for these variations in agreement levels. In Section 3, we delve into the literature on heuristics and develop our theoretical argument regarding the relationship between consistency in the application of heuristic principles and inter-rater agreement on IQ assessments.

3 Theoretical Perspectives: Applying Heuristic Principles in the Assessment of Information Quality

Individuals' inability, and sometimes unwillingness, to apply logic and the rules of probability in making decisions in complex information environments often results in their using “rules of thumb”. Generally speaking, research has characterized these rules of thumb as “heuristics” that individuals apply to make quick decisions or judgments about the object at hand. For example, the great body of the “heuristics and biases” literature that has emerged out of Kahneman and his colleagues' work (Gilovich, Griffin, Kahneman, 2002; Kahneman, 2003; Kahneman, Slovic, & Tversky, 1983; Tversky & Kahneman, 2000) focuses on the individual's use of heuristics for finding adequate, although sometimes imperfect, answers to complex questions. There is a long history of research on heuristics in the behavioral economics and cognitive psychology literature, and, over the last several years, this topic has been the focus of increasing attention in the information systems community. Extrapolating from this literature, one can view heuristics as playing an important role in users' assessment of information quality because they can serve as proxies for more elaborate interactions with content.

Many of the discussions regarding the utility of heuristics in assessing information quality refer to dual-processing models: two widely known dual-processing models are the elaboration-likelihood model (Petty & Cacioppo, 1986; Petty & Wegener, 1999) and the heuristic-systematic model (Chaiken, 1980; Chen & Chaiken, 1999). Both models adopt the working assumption that the amount of time that information receivers devote to evaluating a persuasive message depends on the specific context of use. Two important characteristics of the context of use are users' degree of motivation and the extent to which they can regulate the amount of resources they expend in the process of evaluating the information. For example, in the heuristic-systematic model (Chen & Chaiken, 1999), systematic processing occurs when one needs to fully engage with content, which places a much heavier demand on one's cognitive abilities. Conversely, one can view heuristics as general rules that individuals learn through experience in similar contexts and store in their memory. Given that heuristic processing is much less cognitively demanding, those “who possess little knowledge in the domain” or “individuals who are processing with time constraints” are more likely to use it (Chen & Chaiken, 1999, p. 76).

In the information science and information systems fields, researchers have employed dual-processing models to account for individuals' usage of heuristics for assessing information quality. Metzger (2007) describes heuristics as the default condition when the insignificant consequences will likely result from rendering a poor-quality judgment. As she states, in these situations, "information will be processed or evaluated based on more superficial and less thoughtful criteria" and "decisions will be made on more heuristic judgments of the message or its source (e.g., attractiveness), rather than on message quality" (p. 2087). Sundar (2007, p. 80) defines a heuristic in the context of assessing information quality as "simply a judgment rule (e.g. 'responsiveness is good customer service') that can result in estimations of content quality". Sundar (2007) views the role of these kinds of heuristics as especially important in heterogeneous information environments such as the Web, where there is less consistency in content quality and representation. Cues in the information object elicit heuristics.

Previous studies in the area have identified several cues that individuals use to assess the quality of online content, such as reputation (Metzger, Flanagin, & Medders, 2010) and endorsement (Metzger et al., 2010, Hilligoss & Rieh, 2008). In contrast to a widely held view of heuristics as rules of thumb that are useful but suboptimal and potentially misleading, the view advanced by Gigerenzer and colleagues (Gigerenzer, 2008; Gigerenzer & Gaissmaier, 2011; Gigerenzer & Todd, 1999; Todd & Gigerenzer, 2012) claims that heuristics are simple decision making rules that preclude the need for more traditional, rational decision making strategies. In fact, Gigerenzer claims that, in many instances, one should not consider heuristics as a suboptimal strategy at all because they may yield quicker and better decisions while using fewer cognitive resources (i.e., heuristics are both "fast and frugal"). Although we make no ultimate claim about the comparative correctness of either of these views, we have adopted Gigerenzer's understanding of heuristics for our specific purposes in this paper. This conceptualization recognizes the contextual nature of information use (i.e., it often depends on the task at hand) and stresses the difficulty of making wholly rational decisions in an environment that is extremely heterogeneous as is the case with online information.

The view of heuristics that Gigerenzer (2008) and Gigerenzer and Todd (1999) advance is based on efficiently mapping the task at hand vis-à-vis the variable structures in the information being processed (i.e., "ecological rationality"). Hence, no single set of predetermined heuristics can fit all possibilities of information use. Instead, Gigerenzer and his colleagues (Gigerenzer, 1994; Gigerenzer & Gaissmaier, 2011; Todd and Gigerenzer (2012) propose a set of heuristic building blocks, or ABCs, by which one might construct heuristics in a particular environment. Analyzing cognitive processes in terms of these building blocks "reduc[es] the larger number of heuristics to a smaller number of components, similar to how the number of chemical elements in the periodic table is built from a small number of particles" (Gigerenzer & Gaissmaier, 2011, p. 456). Thus, introducing the notion of heuristic building blocks shifts the focus from identifying the particular heuristics that individuals use across cases to the more general principles on which individuals create specific heuristics in a given use context or environment. In particular, Gigerenzer and Todd (1999) develop three heuristic building blocks and present them as rules. We next describe the three building blocks we used in our investigation.

1) **Search rules are a set of directions** that describe the manner in which one may find relevant, alternative cues or pieces of information (typically through an "active search"). These rules give the search its direction. For example, the "search for cues can be simply random, or in order of some pre-computed criterion related to their usefulness" (Gigerenzer & Todd, 1999, p. 16). Essential to this process is identifying cues in the search environment itself, which, in turn, activate particular heuristics. For example, while assessing the quality of a Wikipedia article, a search rule may direct the information seeker towards the list of references. Here, the information seeker might recognize the name of a respected author, indicating that the referenced source is authoritative and that the citing article has high quality.
2) **Stopping rules** refer to a relatively uncomplicated method for determining when the search should stop. This rule typically concerns the "temporal limitations" of bounded rationality. For example, an individual might terminate a search after encountering two or three relevant cues. To carry on the example from above, once an individual finds that an article has referenced several known authors, the individual searches for no further cues (i.e., the stopping rule states that identifying a few known authors in the list of references is a sufficient criterion for determining the quality of the Wikipedia entry).
3) **Decision rules** enable one to make a choice between alternatives that result from search and stopping rules, or, at the very least, they enable the individual to draw an inference based on the available cues once the individual has stopped searching for cues. A decision rule

indicates a strategy for weighing the accumulated evidence once the search has stopped. Going back to our Wikipedia example, an individual's decision rule may specify that the individual determines the perceived quality of the article based on the number of times it quotes the authoritative source.

Note that decision makers are not aware of these rules; rather, they follow a particular cognitive pattern intuitively (and often unconsciously). Thus, the heuristic building blocks (or rules) represent abstractions that scientists use. Gigerenzer and Brighton (2009) discuss these heuristic principles in the context of a choice between two alternatives. Here, we apply Gigerenzer's framework to a more intricate decision making process: the assessment of an article's quality (along the four IQ dimensions discussed earlier). To further illustrate these ideas, we consider the analogy of a police detective who seeks to determine whether a particular suspect is the murderer. In this analogy, search rules represent the physical paths that the detective takes and the places the detective visits, stopping rules represent the detective's guidelines for having collected sufficient evidence (i.e., cues), and the decision rule specifies a scheme for combining the collected evidence in order to arrive at a decision regarding the suspect's innocence.

In this study, we focus on determining whether there is evidence that inter-rater agreement levels in the assessment of the four IQ dimensions of interest relate to the uniform application of these heuristic principles. We stress that, while the literature on heuristics focuses primarily on an individual's cognitive decision making processes, we extend these ideas and consider the implications for several independent decision makers. Furthermore, our interest extends to determining the effects of consistency in the application of these rules on inter-rater agreement levels in the assessment of the four IQ dimensions mentioned. We conjecture that, given a particular context, high inter-rater agreement in IQ assessment is likely to reflect uniform application of heuristic building blocks and vice versa (i.e., differences in inter-rater assessments are likely to reflect divergence in the application of heuristic building blocks). Figure 1 below illustrates our conjecture.

To summarize, we know that some IQ dimensions are easier to assess than others. Research has shown the use of heuristics to aid in decision making and particularly to facilitate the task of IQ assessment. Here, we investigate whether divergence in the application of heuristic rules may account for differences in the assessments of the various IQ dimensions. Overall, we expand our understanding of the role that cognitive decision making processes play in information quality assessment.

Guided by the three heuristic principles that Gigerenzer and Todd (1999) articulate, we address the following research questions (RQ):

RQ1: To what extent do participants consistently apply heuristic building blocks as they assess an article's quality in terms of accuracy, completeness, objectivity, and representation?

RQ2: What is the degree of inter-rater agreement in the assessment of quality across these four IQ dimensions?

RQ3: Is there an alignment between a) the consistency in the application of heuristic building blocks and b) inter-rater agreement in IQ judgments?

4 Research Method

To address these research questions, we conducted a qualitative and a quantitative study. In the qualitative study, we investigated information seekers' decision making processes when assessing information quality (on the dimensions of accuracy, completeness, objectivity, and representation). We paid particular attention to the extent to which assessors consistently applied the above-mentioned heuristic principles. In the quantitative study, we investigated participants' agreement when assessing IQ (focusing on the same quality dimensions). In both the quantitative and the qualitative portions of the study, we examined two populations of assessors who differed in their information-literacy skills: undergraduate students and university librarians. To this end, we recruited participants from both populations for the qualitative study, whereas, in the quantitative study, we recruited only librarians and relied on the findings of a previous study (Arazy & Kopak, 2011) that employed only student assessors.

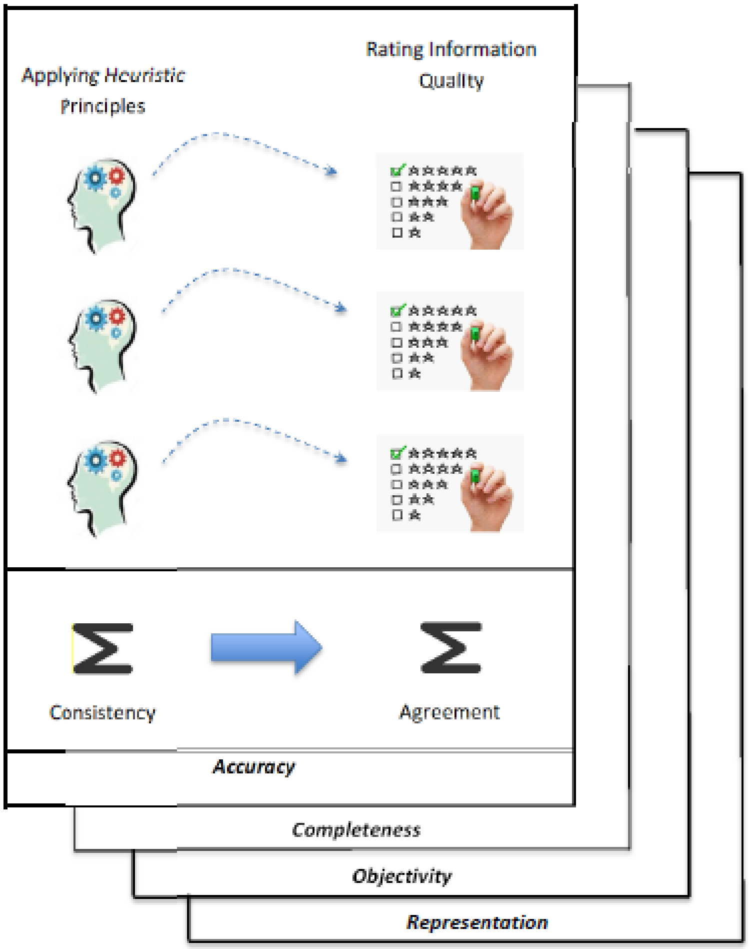


Figure 1. Illustration of Our Theoretical Argument.

To ensure that the participants in all studies had a shared understanding of the IQ dimensions under investigation, we carried out a training session at the beginning of each study in which we discussed the meaning of the four IQ dimensions. We chose Wikipedia articles as the object of assessment given the impact this popular Web resource has had on society (Xu & Zhang, 2013). Moreover, relevant prior studies of information quality investigated the "measurability" of content on Wikipedia (Arazy & Kopak, 2011), which allowed us to draw comparisons. We arrived at a shared understanding of the meaning of various IQ dimensions in the context of Wikipedia as follows: accuracy indicates factual correctness of the data and absence of errors (incorrect information, references to non-authoritative sources, and spelling errors); completeness refers to sufficient coverage of information appropriate for an encyclopedic entry and to the lack of omission of relevant facts (e.g., missing introductory and background information that would help explain the topic's relevance, importance, or its history); objectivity pertains to an impartial view of the topic and to the absence of subjective language, opinions stated as facts, the omission of alternative perspectives or existing controversies, or a deliberate misrepresentation[2]; and representation refers to clarity and ease of understanding at a readership level accessible to the general public (using diagrams when required), rational organization, consistent presentation using a single "voice", and concise formatting.

Next, to ensure that participants were thoroughly familiar with applying these concepts, we asked them to independently analyze the quality of a Wikipedia article of their choosing while paying attention to the four IQ dimensions. Then, a member of the research team provided the participants in training with feedback about applying the IQ criteria to verify there were no ambiguities. To capture data on an exogenous factor that we needed to control for, we began the training session by asking participants to complete a short questionnaire. Using a seven-point Likert scale, participants ranked their reaction to two items that were intended to reflect

[2] Note that participants sometimes used an omission of a relevant fact as evidence for the lack of objectivity.

their disposition towards the quality of content on Wikipedia articles (e.g. “Generally speaking, I believe the quality of information on Wikipedia is high”; “I often use Wikipedia as a source of information”).

4.1 Qualitative Analysis of the Use of Heuristic Principles in IQ Assessment

Our first study focused on revealing the ways in which participants applied heuristic principles when assessing articles' quality. We conducted this study with two populations as assessors: undergraduate students and senior university librarians. For the student assessors, we advertised participation in a second-year undergraduate course; we offered participants a small honorarium. Twelve students signed up to participate in the study. For the session that involved librarians, we sent five senior university librarians a personal email that explained the research objective and invited them to participate. Of the five librarians we approached, three volunteered to take part in our study.

To record the assessors’ thinking processes, we used the think-aloud methodology (Ericsson & Simon, 1993). The task entailed assessing the quality of three Wikipedia articles. To refine the study’s procedure, we first performed a pilot study with four students (other than those participating in the study itself). As part of the pilot, we examined alternative Wikipedia articles and different scenarios. We changed and refined the procedure until we were satisfied that the participants understood the task and that the study’s design would reveal the cognitive processes employed in assessing IQ. The outcome of the pilot study rendered the procedure described below (for a full description of the think-aloud sessions, see Appendix A).

As part of this study's procedure, we asked participants to imagine that, when performing a particular search task, they came across a Wikipedia article (one of the three used in the study) and that their goal was to assess the quality of the article by comparing it against alternative Web resources. We used two information-seeking scenarios: in designing the scenarios, we built on the distinction that Eppler (2006) draws between the objective view of IQ (the extent to which information addresses the information task’s requirements) and the subjective view (the extent to which the information fulfills the user’s expectations). In the scenario corresponding to the objective view, we asked participants to imagine taking part in a research study on text comprehension. They had to evaluate whether a set of comprehension questions that we composed would reflect adequate comprehension of the information in the article. We allocated 30 minutes for this task. In the scenario corresponding to the subjective view, we asked participants to imagine that they wished to impress a friend with their knowledge of a particular topic; they had to gather information in preparation for a meeting with this friend. We allocated 10 minutes for this task. We used this dual scenario research design to indirectly explore the moderating effect of seeker’s motivation (i.e., we designed the 30-minute scenario to motivate participants to spend more effort on the task). To prevent bias due to fatigue, participants began with the more demanding scenario and then proceeded to the second, shorter scenario.

We audio recorded sessions and transcribed the recordings. We based the analysis procedure on verbal protocol analysis principles (Ericsson & Simon, 1993). More specifically, we divided the transcript of each article’s assessment into segments. We classified each segment according to the quality dimension to which it referred and to one of the steps of Gigerenzer’s three-step framework (search direction, stopping rules, and decision rules) (Gigerenzer & Gaissmaier, 2011). Next, we iteratively analyzed each set of classified data (e.g., segments related to search directions for assessing accuracy) analyzed for emergent categories of behavior (e.g., searching for two external sources via a search engine). Next, we scanned the list of categories for each dataset for similarities and merged similar categories (e.g., we merged searching for two external sources in one case with searching for three external sources in another case, which we together called searching for multiple external sources via a search engine).

4.2 Quantitative Analysis of Inter-rater Agreement

To ensure consistency between the two studies, the IQ assessment task in our quantitative study also used Wikipedia articles as the referent. However, the quantitative study focused on whether there was inter-rater agreement in terms of their assessments of the various IQ dimensions. Here, too, we sought to compare the two assessor groups (i.e., students and university librarians). As our baseline, we used Arazy and Kopak’s (2011) study, which used student assessors. We conducted a similar study but employed university librarians as the assessors. Each librarian analyzed all the articles in our set, and we calculated the inter-rater agreement among the three librarians. The documents used in the study included a broad assortment of articles from the English language version of Wikipedia. Specifically, we used the exact same set of Wikipedia articles and the identical versions of the articles that Arazy and Kopak’s (2011) study employed. By using the same articles, we could directly compare the results of the current study (employing librarians) with those of the earlier study (employing students). Appendix B provides additional details regarding the procedure used in the quantitative study.

The librarians assessed the quality of the entire set of Wikipedia articles (printed out on paper in black and white) in random order. We asked them to work independently; to analyze each article carefully; to refer to external sources and compare the content of the article at hand with the content on the same topic found in other sources; to rate the article's quality along the dimensions of accuracy, completeness, objectivity, and representation using a seven-point Likert scale (according to the guidelines developed in the earlier training session); and to count the number of errors (as a proxy for accuracy) and omissions (as a proxy for completeness). When determining their final rating of a particular article, we allowed librarians to refer back to their assessments of other articles in the set.

Once the librarians completed this extensive assessment procedure, we calculated their level of agreement over the entire set of Wikipedia articles for both the rating of each IQ dimensions and the error and omission counts. We opted to use the same metrics as those that Arazy and Kopak (2011) employed so we could make direct comparisons. First, we used the intra-class correlation (ICC) statistic (Haggard, 1958; Landis & Koch, 1977), which is directly analogous to Fleiss' kappa (Fleiss, 1981; Fleiss & Cohen, 1973). Specifically, we used the intra-class agreement metric (range [-1,1]), which emphasizes actual agreement on rating values. Next, to detect cases in which assessments differed yet were in the same direction, we employed the reliability of scale metric (range [-∞, 1]). The reliability of scale signifies ratings' internal consistency and corresponds to the Alpha indicator (which is commonly employed to estimate reliability of instruments). Our method for calculating the statistical significance of differences in interrater reliability followed the approach that Klein, Conn, Smith, and Sorra (2001) and Wong (2008) employed: that is, where one calculates the standard deviation for each of the items (in our case, Wikipedia articles) that multiple assessors rated. In line with Arazy and Kopak (2011), we repeated this calculation for each of the IQ dimensions independently and for the error and omission counts. We then used the assessments' standard deviation as an outcome variable and tested the significance of differences in means using the Mann-Whitney U test (2-sided).

To validate the findings regarding inter-rater agreement for the librarians' quality ratings, we conducted a follow-up analysis once the librarians completed all article ratings. In this second step, the librarians sat together to review one article at a time; they debated and discussed their ratings for each article. A facilitator led a consensus-building process in which the librarians negotiated differences in opinions and worked to reach a consensus on each article's quality, congruent with the Delphi methodology (Linstone & Turoff, 1976). The consensus score was based on a seven-point Likert scale similar to the criteria used in the individual assessment. The consensus-building process was carried out in a quiet office with access to library resources and to the Internet (in case the librarians needed to further research a specific topic). To prevent bias due to fatigue, we spread the meetings out over a month; the librarians met for two hours each time and spent roughly 12 hours in total in attaining a consensus rating along each of the IQ dimensions. To investigate the difficulty of reaching a consensus for each of the articles in our sample, we calculated the differences between librarians' original ratings and the consensus rating. We calculated this distance to consensus metric as follows: assume the rating of a particular assessor i on an article a is ai and assume N assessors (in our case $N = 3$); if $\hat{a}$ represents the consensus rating for article a, then the average distance to consensus for that particular article, $D2C_a = \Sigma|a_i - \hat{a}|N1N/$, and the distance to consensus across the entire set is the average over M articles, $D2CAVG = \Sigma D2CaM1M/$. We calculated this distance to consensus for each of the IQ dimensions.

5 Results

Below we report on the results for the two studies.

5.1 Results for the Qualitative Analysis of the Application of Heuristic

When studying the cognitive processes that underlie the assessment of information quality, we employed Gigerenzer's building blocks framework (Gigerenzer & Todd, 1999) and analyzed participants':

1. **Direction of search**: in particular, we looked at the cues participants attended to and search directions they followed when: a) analyzing the contents of the focal Wikipedia article, b)

consulting external sources (and comparing those to the focal article being assessed), and c) searching their own memory and consulting their prior knowledge of the topic.

2. **Stopping rules***:* the number and kinds of cues that participants used when deciding whether to stop the search.
3. **Decision rules***:* the way in which a participant interpreted the evidence collected (i.e., cues) in making the judgment regarding the article's quality.

Our examination of the think-aloud protocols from the qualitative study revealed participants' cognitive decision making processes when assessing accuracy, completeness, objectivity, and representation. When analyzing these protocols, we estimated the consistency in participants' application of heuristic principles, and we documented the factors that contributed to convergence and those leading to divergence. Given that the results for both students and librarians were similar, we consolidate the findings for both (later we discuss differences between the two assessor groups). Altogether, our study produced 30 article assessments (i.e., 24 by the student assessors (12 participants x 2 articles) and 6 by the librarians (3 participants x 2 articles)). For brevity, we summarize the results related to the assessment of accuracy, completeness, objectivity, and representation below. Appendix C provides the detailed results for the application of heuristic principles (search directions, stopping rules, and decision rules) for each of the IQ dimensions. In the following sections, we describe findings regarding the consistency with which the participants applied the three heuristic rules to each IQ dimension.

5.1.1 The Use of Heuristic Building Blocks When Assessing Accuracy

In terms of search direction, the trajectories participants took when assessing accuracy varied considerably. Broadly speaking, participants applied non-systematic search strategies in terms of both search for cues in the focal Wikipedia article and search in external sources. Furthermore, some participants based their evaluation on their own knowledge and experience (which varied between participants), which further contributed to the inconsistency in the participants' search directions. Stopping rules, too, varied considerably because the portfolio of cues available for assessing accuracy was quite extensive (in all search directions: internal article, external sources, and participants' memory), which led to inconsistencies in the choice of cues that participants considered. In terms of applying decision rules, participants showed a moderate level of consistency, where the factors contributing to divergence in assessment included: 1) the number of cues that participants considered (some based their assessment on very few cues, while others considered a much larger set) and 2) the importance assigned to cues (participants tended to assign higher weight to cues based on their domain knowledge, cues from external sources, or negative cues such as factual errors). On the other hand, the key factor contributing to convergence was the consistency in interpretation: given a specific cue, participants were very consistent in interpreting its implication for accuracy assessment. Overall, the effort and attention that participants exerted in assessing accuracy varied greatly, and we found relatively little consistency in their applying heuristic principles.

5.1.2 The Use of Heuristic Building Blocks When Assessing Completeness

In terms of search directions, participants assessed articles' completeness based on a relatively systematic search strategy, and participants' decision making process was consistent, especially in assessing the focal Wikipedia article (paying particular attention to the table of contents, quickly scanning the document for length and level of detail). When searching external sources, their search direction was also consistent: they focused on a small set of sources (although it is difficult to determine consistency for search directions within those articles). Participants also relied on their personal expectations; thus, they varied in the omissions they identified even when searching in similar directions, which led to divergence. The stopping rules that participants used to terminate the search were highly consistent because the portfolio of cues they used for judging completeness was quite narrow (length and level of detail). When consulting external sources and when relying on their own expectations, participants usually terminated their search after identifying one or two omissions. Decision rules regarding completeness were also quite consistent. Factors that contributed to convergence in assessment included: 1) the small number of cues that participants considered, 2) the importance they assigned to these salient cues, 3) the consistency in how they interpreted cues and their implications for completeness, and 4) the general principle of judging completeness based on the significance of the omission. In contrast, factors contributing to divergence included the reliance on external sources and on participants' expectations (in some of the cases).

5.1.3 The Use of Heuristic Building Blocks When Assessing Objectivity

Objectivity was a more elusive concept, and participants struggled to assess it. Their search directions were quite consistent, and they applied a relatively systematic search strategy: they often assessed only the focal article by scanning it top to bottom in search for opinionated statements. The portfolio of cues they used as stopping rules was quite broad (although not as broad as that employed in assessing accuracy) in that some specific cues grabbed the attention of most participants (namely, use of opinionated language or a structure using “pros-and-cons” sections). However, the texts often buried other evidence for objectivity more deeply, and participants had difficulty identifying them. Even in cases where participants formed an opinion about an article’s objectivity, they often struggled to relate that impression to concrete evidence. The key factor contributing to convergence was the very small number of cues that they considered (most participants relied on a single cue for assessing objectivity, assigning the full weight to that cue). On the other hand, the factors contributing to divergence in assessment included: 1) the difficulty in grounding the impression of objectivity in a concrete fact and 2) the high interpretability of the cues identified. As a result, participants showed relatively high variability in their applying heuristic principles when assessing objectivity.

5.1.4 The Use of Heuristic Building Blocks When Assessing Representation

Participants evaluated representation following a relatively systematic and narrow search strategy, and they showed much consistency in applying heuristics. Search directions were highly consistent (focusing on visual aids and article structure), and the few cases where participants consulted external sources did not add much variance to search direction. Similarly, we observed high consistency in their application of stopping rules (they terminated their search after looking at a few cues, such as images and headers) and decision rules (they consistently interpreted the implications of cues for representation). In summary, four key factors contributed to the high consistency in assessing representation: 1) the focused search direction, 2) the small number of cues employed, 3) the importance assigned to these salient cues, and 4) the straightforward interpretation of these cues.

5.1.5 Summary of Qualitative Results regarding the Use of Heuristic Building Blocks in IQ Assessment

Notwithstanding the challenge of reducing qualitative results into a single score, our research questions required that we compare the various quality dimensions. Table 1 below summarizes the results of the qualitative study: it compares the four information quality dimensions in terms of the extent to which their assessment followed consistent patterns along the three steps of Gigerenzer’s framework (Gigerenzer & Todd, 1999; Gigerenzer et al. 2011).

Table 1. Comparing the Four Information Quality Dimensions in Terms of the Extent to which their Assessment Followed Consistent Patterns along the Three Steps of Gigerenzer’s Framework: Search Direction, Stopping Rules, and Decision Rules (Gigerenzer & Todd, 1999; Gigerenzer et al. 2011)

Consistency in the application of heuristics	Accuracy	Objectivity	Completeness	Representation
Search direction	Low	High	Moderate-high	High
Stopping rules	Low	Moderate	Moderate-high	High
Decision rules	Moderate	Low	Moderate-high	High
Overall consistency	Moderate-low	Moderate	Moderate-high	High

Rank ordering the IQ dimensions, we notice that the participants assessed representation based on the most consistent pattern followed by completeness, objectivity, and accuracy (which recorded the highest divergence in terms of participants’ search directions, stopping rules, and decision rules). It is interesting to note that the dimension that attracted the highest efforts (i.e., accuracy) was also the one where we recorded the lowest consistency (i.e., most noticeable variations) in identifying cues and, consequently, in applying all three heuristic principles. In contrast, representation attracted the least effort and yielded the highest consistency in applying heuristic principles. Assessing both completeness and objectivity called for moderate levels of effort (and time) but for different reasons: completeness because it was relatively simple to assess and cues were easy to spot and interpret and objectivity because it was (at least for some participants) too difficult to operationalize. As a result, when assessing objectivity, assessors did not know which cues to seek, often abandoned the search early on, and based their assessment on unidentified cues.

5.2 Results for Quantitative Study of Inter-rater Agreement

Recall that, in this study, we explore the relationship between the consistency in the application of heuristic principles during IQ assessment and the extent to which IQ assessments demonstrate inter-rater agreement or disagreement. Based on findings from the qualitative study (see above), we could rank order IQ dimensions in terms of the consistency in which participants applied heuristic principles. In this section, we report on the results of our second study in which we analyzed agreement levels for these same IQ dimensions. If the findings demonstrate that the inter-rater agreement on these dimensions mirrors the same order we found in the qualitative study (regarding consistently applying heuristic rules), we may infer that consistently applying heuristic principles is a factor that determines IQ's measurability.

The data collected from the study with the librarians included counts of errors and omissions for each of the Wikipedia articles they assessed in addition to assessors' perceptions regarding the quality of the articles along the various IQ dimensions (indicated on a seven-point Likert scale). We analyzed inter-rater agreement in terms of IQ ratings of the various dimensions and in terms of the error (a proxy for accuracy) and omission (a proxy for completeness) counts. Generally speaking, we found inter-rater agreement levels to be moderate. Landis and Koch (1977) provide a scale for interpreting the Kappa inter-rater value. Fleiss (1981) and Fleiss and Cohen (1973) interpret ICC values in a similar way to Landis and Koch. The Landis and Koch (1977) scale suggests that values below 0.20 represent poor agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, and 0.61-0.80 substantial agreement. However, note that this scale represents a generalization, and agreement levels depend on the number of categories (Sim & Wright, 2005). Thus, the low ICC results for the rating of articles' quality (0.07-0.24) could arise from the broad range used for rating the IQ (i.e., 7). ICC values for the error and omission count were higher (0.31-0.53). Also, internal consistency as measured through scale reliability was higher than ICC with values in the range of 0.5-0.8 (see Tables 2 and 3 below), which represents moderate-high agreement levels.

Table 2. Inter-rater Agreement Results for the Various Constructs

	Intra-class agreement (ICC)	Reliability of scale (unbiased)
Accuracy	0.157	0.562
Completeness	0.236	0.547
Objectivity	0.141	0.615
Representation	0.218	0.602

When analyzing the differences in inter-rater reliability between the various quality dimensions, we notice that, in terms of ICC, the participants attained highest agreement level for completeness followed by representation, accuracy, and objectivity. However, scale reliability results (see Table 2) revealed a somewhat different story: the results show relatively high scale-reliability for all quality dimensions, though the score for objectivity (which was low on ICC) was highest. Another effect that was inconsistent with ICC findings was the low scale-reliability score for completeness (which scored highest on ICC). Such differences can result, for instance, when all assessors agree on which articles have high quality and which have low quality except for one assessor who is more stringent (or, conversely, more lenient) than the others (e.g., the one assessor consistently rates all articles as lower in quality but in the same relative order of quality as the other assessors). This situation would result in low ICC and high scale reliability (e.g., as we observed for objectivity).

We performed a similar analysis of agreement for the error and omission counts, and Table 3 presents the results. ICC values were substantially higher for the omission count (0.53 compared to 0.31 for the error count). This result corroborates the findings from the IQ dimension assessments, where the ICC value for completeness (which corresponds to the omission count) was substantially higher than the ICC value for accuracy (which corresponds to error count). Interestingly, in contrast to what we observed for the IQ dimension assessments, in the errors and omission count, the pattern of results obtained using the ICC was consistent with that obtained using the reliability of scale measure.

Table 3. Inter-rater Agreement Results for the Errors and Omissions Count

	Intra-class agreement (ICC)	Reliability of scale (unbiased)
Error count	0.305	0.578
Omission count	0.527	0.775

The findings from the consensus-building portion of the study (see Table 4) are consistent with the results of the previous analyses (namely, the ICC results for the quality ratings and the results of both agreement metrics in the error and omission count) and indicate that reaching an agreement—as measured by the differences between librarians' original ratings and their consensus ratings—was easiest for completeness followed by representation, accuracy, and objectivity (with average distances of 0.48, 0.64, 0.72, and 0.89, respectively). Using the Mann-Whitney U test, we found that differences in distance to consensus between pairs of IQ dimensions were all statistically significant ($p < .01$) except for the difference between accuracy and representation[3].

Table 4. Distance to Consensus Results for the Various Constructs

	Average	Standard deviation
Accuracy	0.72	0.53
Completeness	0.48	0.42
Objectivity	17.7	11.5
Representation	0.89	0.43
Factor 5	0.64	0.51

Summarizing the study of inter-rater agreement in IQ assessment, Table 5 uses a categorical scale to consolidate the results from the various analyses (quality ratings, counts of errors and omissions, and the distance to consensus measure[4]). In sum, the data in Table 5 indicate that the participants obtained the highest agreement levels for representation and completeness followed by objectivity and, with the lowest agreement levels, accuracy.

Table 5. Inter-rater Agreement Based on the Various Measures

Measure of agreement		Accuracy	Objectivity	Completeness	Representation
IQ perception rating	ICC	Moderate-low	Low	High	Moderate-high
	Scale reliability	Moderate-low	High	Low	Moderate-high
Error & omission count	ICC	Low	NA	High	NA
	Scale reliability	Low	NA	High	NA
Proximity to consensus		Moderate	Low	High	Moderate-high
Overall agreement		Moderate-low	Moderate	Moderate-High	Moderate-high

Comparing the results of librarians' quality ratings to the findings from Arazy and Kopak (2011) (who employed student assessors) further corroborates our results. Namely, the IQ dimensions' rank order is almost identical between the two studies such that dimensions that yielded high agreement in our study (completeness and representation) also yielded relatively high inter-rater agreement in Arazy and Kopak's (2011) study (and similarly for the low-consistency dimensions, accuracy and objectivity). However, interestingly, the overall agreement levels in our study (with librarians as assessors) were significantly higher than those that Arazy and Kopak (2011) report.

Finally, comparing the results of our two studies (i.e., the qualitative analyses of assessors' cognitive decision making processes and the quantitative study of inter-rater agreement in IQ assessment) reveals

[3] With the p values of .001, .003, .001, .008, and .001, respectively, for the pairs accuracy-completeness, accuracy-objectivity, completeness-objectivity, completeness–representation, and objectivity-representation.
[4] For consistency with the other measures, the table transposes this into "similarity to consensus".

substantial alignment. In particular, the ordering of inter-rater agreement levels mirrored the rank ordering of IQ dimensions in terms of consistency in applying heuristic principles.

6 Discussion

In this research, we investigated whether certain cognitive decision making processes could explain differences in assessors' ratings of various IQ dimensions. We conducted two studies: a qualitative study and a quantitative study. In the qualitative study, we investigated the cognitive process by which university students and librarians assessed the quality of three Wikipedia articles along the dimensions of accuracy, completeness, objectivity, and representation. We used the concept of building blocks that Gigerenzer and Todd (1999) propose as bottom-up strategy in lieu of a formal top-down framework to identify the existence of a heuristic across all information environments. Through the behaviors and utterances of our participants, we observed that they used three types of building blocks when assessing four IQ dimensions. We found that participants converged in their application of heuristic principles when assessing representation and completeness but diverged in their application when they assessed accuracy and objectivity. In the quantitative study, we recruited university librarians as participants to judge the quality of a larger set of Wikipedia articles. Adopting the procedure from Arazy and Kopak (2011) for each Wikipedia article in our set, the three librarians produced quality assessments along the various dimensions and a count of the number of errors (which corresponds to accuracy) and omissions (which corresponds to completeness). Our analyses measured inter-rater agreement in the judgments of accuracy, completeness, objectivity, and representation and in the counts of errors and omissions. In addition, we asked the three librarians to try and reach an agreement on the quality of each article along the four dimensions, and we measured the average distance to consensus for each of the IQ dimensions. Put together, the quantitative analyses showed that some dimensions (completeness and representation) consistently yielded higher inter-rater agreement than others (accuracy and objectivity). We then compared the findings from the qualitative and quantitative study and found high correspondence between the consistency in the application of heuristic principles and inter-rater agreement in IQ judgments.

6.1 Applying Heuristics Theory to the Study of Information Quality

The primary contribution of our study is in associating inter-rater agreement in IQ judgments with the cognitive processes that individuals use to assess information. Namely, our findings indicate that the IQ dimensions in which we observed participants consistently apply heuristic principles (completeness and representation) also yielded relatively high inter-rater agreement levels; similarly, the IQ dimensions in which we observed participants less consistently apply heuristic principles (accuracy and objectivity) yielded relatively low inter-rater agreement levels. To the best of our knowledge, no prior study has empirically demonstrated this linkage between the use of decision making processes and the resulting IQ agreement scores.

How could inconsistencies in the application of heuristic principles lead to disagreements between independent assessors of information quality? We offer three possible explanations linked to the three heuristic rules in Gigerenzer's framework (Gigerenzer & Todd, 1999; Gigerenzer et al. 2011). First, variation in the type of search rules that assessors select suggests that assessors follow different search paths; thus, the types of evidence (or cues) they encounter are bound to be different, which, in turn, leads to different decisions. For example, when assessing accuracy, one assessor might rely heavily on external sources, another might rely on personal knowledge, and a third might consider the logic of the argument presented in the text. Hence, we can expect divergence in their perceptions of the article's quality, which will result in disagreement in evaluations. Likewise, consistency in terms of search directions in the focal article (as in the case of representation) may contribute to agreements in IQ judgments. Second, variations in stopping rules may eventually lead to disagreements in judgment. For example, our participants used a variety of stopping rules to assess accuracy (e.g., some participants relied on a single error to judge the article as inaccurate, while others sought multiple indicators). In contrast, when assessing representation, participants consistently terminated their analysis after skimming for graphics, information boxes, and headers. Finally, participants who applied similar decision rules translated into agreements in IQ judgment. In some IQ dimensions, participants consistently applied decision rules. For example, in assessing representation, fluent language or use of pictures consistently evoked positive perceptions of quality. Similarly, participants consistently associated long texts with perceptions of high completeness. In contrast, we observed much variation in participants' decisions regarding articles' objectivity because cues were ambiguous and their meaning subject to interpretation.

Hilligoss and Rieh (2008) describe three classes of short-hand decision making devices for assessing information quality, among them “general rules of thumb” that scholars have broadly applied to a variety of objects and circumstances[5]. These rules of thumb are closely related to the results of the heuristic principles we used in our analysis. For example, a general rule of thumb for completeness would be: “look for the presence of a bibliography and check the length of content, stop looking, and make an assessment based on the extent of these two cues”. Similarly, one may judge representation based on consistency in structure and page design (Flanagin & Metzger, 2007). These kinds of general decision making devices are largely topic independent and do not require specific domain knowledge on the user’s part of the user; thus, independent assessors tend to apply them consistently, which, in turn, leads to higher agreement between them. Considering the particular genre of writing, one could more easily judge an article as “complete” if it has a certain length and is well documented through including relevant references regardless of the topic. In contrast, a class of general rules of thumb might not be readily available when it comes to judging accuracy and objectivity because their assessment may require one to closely read the content. For example, applying different decision rules to determine the credibility of an external source resulted in variations in how our participants assessed objectivity: some interpreted the article's similarity to an external source as undermining its credibility, while others interpreted the similarity as enhancing the article's credibility (especially in cases when the external source was a not-for-profit institution).

Our results regarding the cognitive processes used in IQ assessment inform the cognitive psychology literature on the use of heuristics in decision making. Namely, our findings add validation to Gigerenzer and Todd’s (1999) building blocks framework and demonstrate the usefulness of this conceptualization. An important contribution of our study is our applying Gigerenzer and Todd’s framework to the study of information quality. By using this framework in this particular context, we could organize the recorded cognitive trajectories based on (common or diverging) search directions, stopping rules, and decision rules. Moreover, by focusing on the three building blocks, we could distinguish between the processes employed for analyzing various IQ dimensions.

We stress that we were not able to straightforwardly apply a cognitive theory to this specific context; rather, we needed to reconcile conceptualizations of two separate scholarly fields. Furthermore, Gigerenzer and his colleagues (Gigerenzer & Brighton, 2009) developed their ideas in the context of a simple cognitive process: a choice between two alternatives. In our study, we applied these ideas to the much more complex process of IQ assessment. We note that prior works on IQ assessment that reference heuristics (Metzger & Flanagin, 2013) use the term rather loosely (e.g., mixing the notions of cues and heuristics) and do not make a direct linkage to theoretical frameworks of decision making. We further extend the building blocks framework to consider the implications for a group of decision makers, whereas Gigerenzer's group (and, more broadly, decision making theories) analyzed cognitive processes at the individual level. However, our conceptualization emphasizes consistency (or divergence) in the application of particular heuristic principles among a group of independent IQ assessors.

Further, with our qualitative study, we demonstrate that a distinct set of cognitive decision making processes (namely, the three heuristic principles) is associated with each of the IQ dimensions. While prior studies on the use of heuristics in IQ assessment (Lim, 2013; Metzger et al., 2010; Rowley & Johnson, 2013) examined the high-level construct of information quality (and credibility), they did not distinguish between the particular heuristics used (e.g., those employed in assessing accuracy and those employed in assessing completeness). In particular, Metzger’s and Flanagin’s (2013) conceptualization of IQ assessment was based on the notion of cognitive heuristics and did not distinguish between different IQ dimensions. Hence, we not only employ Gigerenzer’s (Gigerenzer & Todd, 1999; Gigerenzer et al. 2011) ideas about the IQ assessment but also show that distinct searching, stopping, and decision rules are associated with the assessment of different IQ dimensions. Thus, our findings suggest that research on information quality (or credibility) should frame discussions about IQ assessment at the level of the particular IQ dimension (e.g., accuracy) rather than at the level of the composite construct of information quality.

6.2 Implications for Studies on the Measurability of Information Quality

Importantly, in our quantitative study, we identified substantial differences between inter-rater reliability scores for the different quality dimensions such that we found lower agreement in the ratings of some indicators compared to others. Using the inter-rater reliability (IRR) metric, we found that completeness

[5] According to Hilligoss and Rieh (2008), the two other decision making devices for assessing information credibility include the “construct” and “interaction” levels.

and representation yielded higher agreement levels than accuracy and objectivity. We observed similar results when we analyzed agreement levels in the counts of errors and omissions. Our study introduces a novel method for estimating multi-rater agreement through using the Delphi methodology (Linstone & Turoff, 1976) and measures the distance between assessors' original ratings of IQ and the later consensus rating. Comparing the distance to consensus and the IRR results from our first study, we found an identical ranking in IQ dimensions measurability.

Notwithstanding the overall higher agreement levels in our study of librarians, we found a striking resemblance between the librarians' and the students' assessment of Wikipedia articles in (Arazy & Kopak, 2011) in terms of the agreement ranking: in both studies completeness yielded the highest agreement levels followed by representation, accuracy, and objectivity. Moreover, our results indicate that, in a particular context, the rank order of agreement between IQ dimensions is stable across agreement metrics. We do not argue this exact rank order of agreement would generalize to every IQ assessment context. In fact, we believe that the measurability of information quality may depend on media type and task context. For example, we expect that, in analyzing the content of blogs, assessors would find it difficult to agree on completeness given the lack of clear expectations.

Carefully analyzing our results sheds some light on the moderating factors that affect the observed patterns. In particular, we studied the ways in which assessors' information literacy skills and their motivation in the assessment task affected both how they applied heuristic principles and inter-rater agreement levels in IQ judgments. First, we attend to the effect of information literacy. In our study, we did not directly estimate assessors' information literacy levels; however, there are clear literacy differences between ordinary information users (undergraduate students) and information professionals (senior university librarians) (McDowell, 2002). In the qualitative study, we observed few differences between the assessor populations: both followed similar search directions that were anchored on the same cues and employed comparable decision rules. The primary difference was that librarians tended to read the article more carefully and paid more attention to the list of references at the bottom of the Wikipedia article (and followed these references in their external search). Also, we noticed small variations in terminology and in the attention paid to the task (librarians used a more professional terminology and commented on the scenarios). Overall, we found similar patterns across the two assessor groups in how they applied heuristic principles for the different IQ dimensions. Thus, for example, both students and librarians were more consistent in applying heuristic principles when assessing representation and completeness and less consistent in assessing accuracy and objectivity. Interestingly, we observed similar patterns in the quantitative study. Comparing the agreement levels of the librarian assessors in our study to the agreement levels of student assessors in Arazy and Kopak (2011) sheds light on how information literacy skills affect agreement levels in IQ assessment. Our analysis shows that, although inter-rater agreement levels (in terms of both ICC and scale reliability) were higher for librarians, the rank order of IQ dimensions was consistent. These results indicate that information literacy does not affect the relationship between the application of heuristic principles in IQ assessment and inter-rater agreement. This conclusion is in line with Lim (2013), who found that the effect of peripheral cues on perceived credibility was similar for users with varying degrees of knowledge.

We studied motivation's potential moderating effect by comparing a low-effort scenario with a higher-effort scenario. Simon (1980) explains that the use of heuristics depends on the task environment, and prior studies have shown that individuals activate heuristics by default as a first choice of reasoning, whereas they adopt more exhaustive decision making strategies (i.e., "rational") optionally with respect to particular conditions, such as decision maker's motivation (Evans, 2006; Kahneman & Frederick, 2002). Dual-processing models predict that, when individuals have low motivation and/or ability, they will process or evaluate information on more superficial and less thoughtful criteria[6]. In these situations, they will make decisions based on more heuristic judgments of the message or its source. As the motivation to engage with the content and the user's ability to make judgments about the content increases, the likelihood individuals will use systematic decision making instead of heuristics also increases. When analyzing differences between scenarios in the qualitative study in terms of the directions participants searched and the stopping and decision rules they employed, we noticed some variations. Namely, participants had a higher tendency to compare the focal article to external sources in the case of the long scenario (13 out of 15 cases) compared to the short scenario (9 out of 15 cases). For example, one participant commented that: "for the purpose of meeting a friend [the short scenario], this is enough". We found these differences

[6] With this said, note that Gigerenzer (2008) would disagree that applying heuristics is always inferior to more rational approaches.

primarily when participants assessed accuracy and completeness. We observed only small differences between scenarios when participants assessed objectivity, likely because they struggled to assess this particular IQ dimension because they were not sure which cues or anchors to look for (and, thus, also spent relatively little time on the more demanding scenario). We observed no differences for representation, and, in both scenarios, assessment was immediate. Overall, we found little evidence to suggest that the differences between the various IQ dimensions depended on the task, and, in both tasks, we recorded high consistency in participants' applying heuristic principles for representation and completeness and lower consistency for accuracy and objectivity.

6.3 Limitations and Future Research

This study has several limitations one should consider when interpreting the results. The first concerns the qualitative study's assessors' cognitive processes. While the think-aloud method focuses on allowing one to follow subjects' thoughts, we had no certainty that the participants indeed reported their full thinking processes. The number of participants in this study presents an additional drawback in that it limits the generalizability of our conclusions. Nonetheless, small samples are common in such qualitative studies given the investment required from each participant and the effort involved in analyzing the rich data. Our quantitative study of inter-rater agreement also has a few limitations. First, we based the inter-rater agreement on only three assessors. For comparison, we needed to match the task in Arazy and Kopak's (2011) earlier study (i.e., same set of articles; three assessors per article) while addressing its methodological shortcomings and ensuring that the same set of raters analyzed all information objects. As a result, we assigned each assessor a complex and time-consuming task: the detailed examination of 98 Wikipedia articles, which involved using other resources to research the articles' topic. This task required approximately 100 hours (including the follow-up consensus-building study). We also acknowledge that there may be some concerns regarding potential biases in our sample of assessors (e.g., assessors' expertise or background), and one should exercise caution when generalizing the results to other highly information-literate populations. That being said, the design of our quantitative study addresses these concerns to a large extent. We ensured that the set of Wikipedia articles we used covered a wide spectrum of topics in Wikipedia; thus, it is unlikely that one librarian had substantially more domain expertise than the others. Furthermore, to address the issue that one rater applied different standards from the others, we used (in addition to ICC) the reliability of scale metric, which examines the correlation between the assessors' ratings rather than the agreement of the actual values. Thus, if one of the librarians consistently applied either stricter or more lenient standards, it would not affect this metric. In the future, we hope to repeat our study with a larger sample of assessors and possibly assessors with different skill sets (e.g., domain experts) and directly measure and control for exogenous factors such as assessors' cognitive or demographic traits (e.g., age, computer self-efficacy, information literacy, domain knowledge).

A limitation of both our studies is that they investigated one information resource: Wikipedia. Thus, Wikipedia's distinct characteristics or our assessors' predispositions toward it may have affected our findings. For example, the standard article structure of Wikipedia and Wikipedia's ubiquity acts to set readers' expectations and helps them make judgments on completeness (assessing the completeness of a blog's postings, for example, would be much more difficult). Similarly, Wikipedia's standard formatting may affect assessments of representation. We plan to repeat this study on alternative information sources, including both user generated content such as Yahoo! Answers and traditional resources (e.g., Consumer Reports), and pay particular attention to the potentially moderating role of an articles' quality. Also, the set of quality dimensions we employed in our study provides only a partial representation of this multi-dimensional construct, and we propose that future studies expand our investigation to additional quality dimensions, (e.g., timeliness, understandability).

We conclude that information quality is an elusive construct that is difficult to measure, and users' quality assessments are subjective and depend on the manner in which they apply the three heuristic building blocks: search direction, stopping rules, and decision rules, which makes it difficult for multiple assessors to reach an agreement on a resource's quality. Our study provides some novel insights regarding the effects of assessors' cognitive processes (and, in particular, the application of heuristic principles) on information quality judgments. Nevertheless, additional research needs to validate our findings in alternative settings, expand the scope of investigation, and explore the role of additional factors that affect variations in inter-rater agreement levels.

7 Conclusion

We conclude our discussion with implications of our findings to research in information systems and for practice.

7.1 Implications for Information Systems Research

Our work suggests that IS scholars need to recognize the difficulty of measuring IQ and remember that it is not a unidimensional construct and that some dimensions lend themselves more readily to consistent judgment. Furthermore, researchers should consider assessors' tendency to rely on heuristics; as our findings suggest, how individuals apply heuristic principles is a key factor in determining their ability to reach an agreement. Note, however, that the understanding that higher inter-rater agreement levels stem—at least in part—from a consistent application of heuristic principles suggests that even assessments that reflect high agreement levels may, in fact, suffer from biases. For example, individuals may consistently interpret a lengthy description as complete information even when the description misses important details. Such biases are more likely in cases where assessors are limited in their information literacy or domain knowledge. Based on these significant concerns regarding the reliability of IQ measures, we suggest that studies that employ IQ as either a dependent or independent variable should be more careful in measuring this construct. Moreover, future studies should seek a reliable method for determining the validity of the IQ measurements.

7.2 Implications for Practice

A practical recommendation for information users is to take greater care in judging quality and in accepting others' quality ratings given that assessing information quality invites biases. This recommendation is especially applicable to user-generated content such as social news and product review websites (Agichtein, Castillo, Donato, Gionis, & Mishne, 2008; Harper, Raban, Rafaeli, & Konstan, 2008). Notwithstanding the advantages of heuristics in allowing quick and cognitively efficient judgments, consumers of information should recognize that these heuristics might provide only a limited indication of the overall quality of the object.

Another practical recommendation concerns the providers of Web services that produce information quality metrics for their published content. As Arazy and Kopak (2011) suggest, providers that rely on users' assessments should seek to employ a large number of users, particularly for the IQ dimensions of accuracy and objectivity. We suggest that one should consider users' expertise in the particular domain of interest (and possibly give extra weight to the assessments of more knowledgeable users). A broader implication of our results is the need to triangulate users' ratings with some external, less-biased sources. This point is particularly relevant for IQ dimensions that yield low agreement levels (but even high agreement levels cannot ensure reliability because agreement can result from individuals' consistently using a heuristic). Finally, a practical implication for educators of information literacy is to be more conscious of the limitations of heuristics for analyzing information and to place greater emphasis on developing assessment skills and techniques.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the International Conference on Web Search and Data Mining* (pp. 183-194).

Arazy, O., & Kopak, R. (2011). On the measurability of information quality. *Journal of the American Society for Information Science and Technology, 62*(1), 89-99.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom.* New Haven, CT: Yale University Press.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752-766.

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology.* New York: The Guildford Press.

Chesney, T. (2006). An empirical examination of Wikipedia's credibility. *First Monday, 11*(11).

Cicchetti, D. V., & Heavens, R. (1981). A computer program for determining the significance of the difference between pairs of independently derived values of kappa or weighted kappa. *Educational and Psychological Measurement*, *41*(1), 189-193.

Eppler, M. J. (2006). *Managing information quality: Increasing the value of information in knowledge-intensive products and processes* (2nd ed.). Berlin: Springer.

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

Evans, J. S. B. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13*(3), 378-395.

Eysenbach, G., Powell, J., Kuss, O., & Sa, E. R. (2002). Empirical studies assessing the quality of health information for consumers on the World Wide Web. *Journal of the American Medical Association*, *287*(20), 2691-2700.

Fallis, D. (2008). Toward an epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology*, *59*(10), 1662-1674.

Fink-Shamit, N., & Bar-Ilan, J. (2008). Information quality assessment on the Web—an expression of behavior. *Information Research, 13*(4).

Flanagin, A. J., & Metzger, M. J. (2003). The perceived credibility of personal web page information as influenced by the sex of the source. *Computers in Human Behavior, 19*(6), 683-701.

Flanagin, A. J., & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of Web-based information. *New Media & Society, 9*(2), 319-342.

Fleiss, J. L. (1981). The measurement of interrater agreement. In J. L. Fleiss (Ed.), *Statistical methods for rates and proportions* (pp. 212-236). New York: John Wiley.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3), 613-619.

Fleiss, J. L., Cohen, J., & Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*(5), 323-327.

Gagliardi, A., & Jadad, A. R. (2002). Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination. *BMJ, 324*(9), 569-573.

Gigerenzer, G. (1994). Where do new ideas come from? In M. A. Boden (Ed.), *Dimensions of creativity* (pp. 53-74). Cambridge, MA: MIT Press.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford, UK: Oxford University Press.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451-482.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 3-34). Oxford, UK: Oxford University Press.

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior*. New York: Oxford University Press.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature, 438*(7070), 900-901.

Gilovich, T., Griffin, D. W., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgement.* New York: Cambridge University Press.

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance*. New York, NY: Dryden Press.

Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A. (2008). Predictors of answer quality in online Q&A sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 865-874).

Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management, 44*(4), 1467-1484.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, *93(*5), 1449-1475.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment.* New York: Cambridge University Press.

Kahneman, D., Slovic, P., & Tversky, A. (1983). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Kim, P., Eng, T. R., Deering, M. J., & Maxfield, A. (1999). Published criteria for evaluating health related web sites: Review. *BMJ,* 318(7184), 647-649.

Klein, K. J., Conn, A. B., Smith, D. B., & Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology, 86*(1), 3-16.

Knight, S., & Burn, J. (2005). Developing a framework for assessing information quality on the World Wide Web. *Informing Science*, *8*, 159-172.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lankes, R. D. (2008). Trusting the internet: New approaches to credibility tools. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 101-122). Boston, MA: MIT Press.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815-852.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management, 40*(2), 133-146.

Lim, S. (2009). How and why do college students use Wikipedia? *Journal of the American Society for Information Science and Technology*, *60(*11), 2189-2202.

Lim, S. (2013). College students' credibility judgments and heuristics concerning Wikipedia. *Information Processing & Management*, *49*(2), 405-419.

Linstone, H. A., & Turoff, M. (1976). *The Delphi method: Techniques and applications.* New York: Addison-Wesley.

Liu, Z. (2004). Perceptions of credibility of scholarly information on the Web. *Information Processing & Management, 40*(6), 1027-1038.

Lukyanenko, R., Parsons, J., & Wiersma, Y. (2014). The IQ of the crowd: Understanding and improving information quality in structured user-generated content. *Information Systems Research, 25*(4), 669-689.

Luyt, B., Aaron, T. C. H., Thian, L. H., & Hong, C. K. (2008). Improving Wikipedia's accuracy: Is edit age a solution? *Journal of the American Society for Information Science and Technology, 59*(2), 318-330.

McDowell, L. (2002). Electronic information resources in undergraduate education: An exploratory study of opportunities for student learning and independence. *British Journal of Educational Technology*, 33(3), 255-266.

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, *58*(13), 2078-2091.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics, 59*, 210-220.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, *60*(3), 413-439.

Michnik, J., & Lo, M. C. (2009). The assessment of the information quality with the aid of multiple criteria analysis. *European Journal of Operational Research, 195*(3), 850-856.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation*, *7*(10)

Nurse, J. R. C., Creese, S., Goldsmith, M., & Lamberts, K. (2011). Trustworthy and effective communication of cybersecurity risks: A review. In *Proceedings of The 1st Workshop on Socio-Technical Aspects in Security and Trust* (pp. 60-68).

Oakleaf, M. (2009). Using rubrics to assess information literacy: An examination of methodology and interrater reliability. *Journal of the American Society for Information Science and Technology, 60*(5), 969-983.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 123-205). New York: Academic Press.

Petty, R. E., & Wegener, D. T. (1999). The elaboration likelihood model: Current status and controversies. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology.* New York: The Guildford Press.

Rains, S. A., & Karmikel, C. D. (2009). Health information-seeking and perceptions of website credibility: Examining Web-use orientation, message characteristics, and structural features of websites. *Computers in Human Behavior*, *25*(2), 544-553.

Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology, 41*(1), 307-364.

Rowley, J., & Johnson, F. (2013). Understanding trust formation in digital information sources: The case of Wikipedia. *Journal of Information Science*, *39*(4), 494-508.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257-268.

Simon, H. A. (1980). Cognitive science: The newest science of the artificial. *Cognitive Science, 4*(1), 33-46.

Stanford, J., Tauber, E., Fogg, B., & Marable, L. (2002). Experts vs. online consumers: A comparative credibility study of health and finance web sites. *Consumer WebWatch.*

Stvilia, B., Twidale, M., Smith, L., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*(6), 983-1001.

Sundar, S. S. (2007). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger (Ed.), *Digital media, youth, and credibility* (pp. 73-100). Cambridge, MA: MIT Press.

Taylor, R. S., & Voigt, M. J. (1986). *Value added processes in information systems.* Westport, CT: Greenwood Publishing Group.

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. New York: Oxford University Press.

Tversky, A., & Kahneman, D. (2000). *Choices, values, and frames*. Russell Sage Foundation.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-33.

West, K., & Williamson, J. (2009). Wikipedia: Friend or foe? *Reference Services Review, 37*(3), 260-271.

Wong, S. S. (2008). Task knowledge overlap and knowledge variety: the role of advice network structures and impact on group effectiveness. *Journal of Organizational Behavior*, *29*(5), 591-614.

Wong, S. S. (2008). Task knowledge overlap and knowledge variety: The role of advice network structures and impact on group effectiveness. *Journal of Organizational Behavior, 29*(5), 591-614.

Xu, S. X., & Zhang, X. (2013). Impact of Wikipedia on market information environment: Evidence on management disclosure and investor reaction. *MIS Quarterly, 37*(4), 1043-1068.

Yaari, E., Baruchson-Arbib, S., & Bar-Ilan, J. (2011). Information quality assessment of community generated content: A user study of Wikipedia. *Journal of Information Science, 37*(5), 487-498.

Appendix A: Additional Details Regarding the Method for the Qualitative Analysis of the Application of Heuristic Principles in IQ Assessment

This appendix provides details regarding the qualitative study's methodology.

For both student and librarian assessors, we administered the think-aloud sessions with each participant as follows. The interviewer (one of the researchers) first discussed with the participants the various IQ dimensions and made sure they had a clear understanding of the constructs' definitions. Next, to ensure participants' familiarity with the standards of Wikipedia articles (in terms of structure, style, and format), we asked each participant to browse through a few Wikipedia articles of their own choice while keeping the IQ dimensions in mind. Following these introductory steps, we introduced the participants to the two assessment tasks and gave them instructions for the assessment procedure. Specifically, for each of the two scenarios, we assigned participants to a particular article and then asked them to evaluate its quality along the relevant IQ dimensions. We instructed the participants to think aloud as they assessed the articles' quality, the procedure that Ericsson and Simon (1993) recommend. Once the study's set-up was complete, we reminded the participants that we would record the session and that we had activated the recording device.

Once underway, we presented the participants with the first scenario and pointed them to the Wikipedia article they had to assess. The participants received a printed copy of the instructions and assessment sheets and used a computer to read the focal Wikipedia article and to search online for and read any additional material they found relevant for making their assessment. As soon as we presented the article, we asked the participants to answer a question about their knowledge of the article's topic. Then, we asked the participants to assess the quality of content of the Wikipedia article along the various IQ dimensions of interest (by reading the article and comparing it with other Web resources). After completing the first scenario, the participant proceeded to the second while following a similar protocol (we alternated the scenario order between participants). During this entire procedure, if participants forgot to think aloud, we gently reminded them to do. When completing the second task, the interviewer guided an open discussion with the participant covering both scenarios. The interviewer asked the participants to reflect on the way in which they assessed each of the IQ dimensions, the difficulty in making a judgment about the quality of the article, and the cognitive decision making process they used. The duration of these think-aloud sessions varied between 55 and 90 minutes.

Appendix B: Additional Details Regarding the Method for the Quantitative Study of Inter-rater Agreement in IQ Judgments

This appendix provides details regarding the quantitative study's methodology.

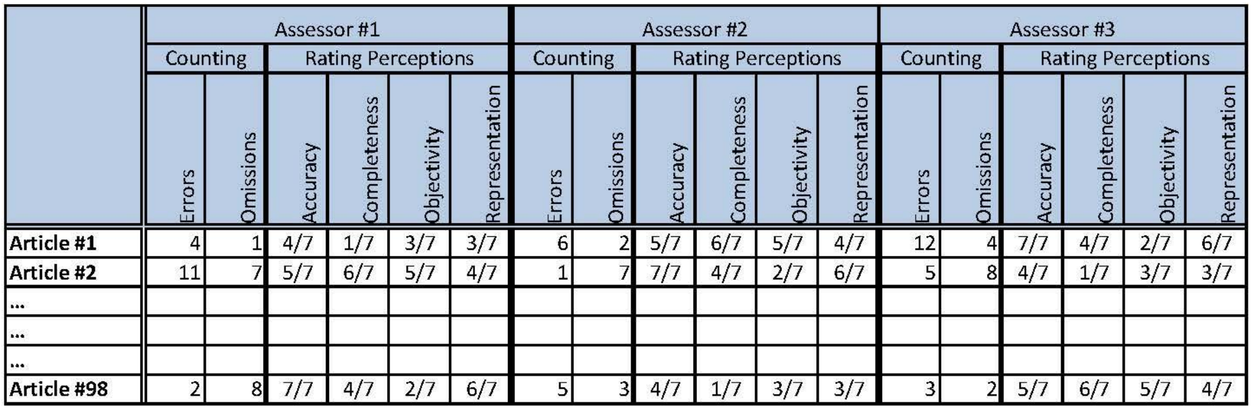
We based our procedure for analyzing the measurability of IQ on the one Arazy and Kopak (2011) used. In that study, the authors randomly assigned students into 10 assessor groups of three participants per group (with some overlap between them). Each student assessed only a subset of the articles such that overall, three to six students assessed each article. The authors calculated inter-rater agreement for each group. The agreement values were based on the average of all groups.

In our study, we employed three librarians as assessors, and each librarian analyzed all of the articles in our set. We calculated inter-rater agreement among the three librarians. Our assessors assessed the same documents the assessors in Arazy and Kopak (2011) did. As such, we could make direct comparisons to the results of that study. The set of 98 Wikipedia articles included articles of 200-3500 words (to eliminate stubs and exceptionally long outliers) with an equal representation of six topical categories: 1) culture, art, and religion; 2) math, science, and technology, 3) geography and places, 4) people and self, 5) society, and 6) history and events. By having subsets of articles from different topical categories, we could compare agreement levels between categories. Given that Wikipedia articles are in a state of continuous flux, we ensured that the articles used were the exact same version of articles as in Arazy and Kopak (2011).

The procedure for assessing the articles' quality comprised several steps. After the training session in which the librarians reached a shared understanding of the various IQ dimensions, we worked to develop clear criteria for judging articles' quality along the four IQ dimensions. The librarians independently analyzed six different Wikipedia articles from various topical categories (these were articles not included in the study set described above). The analysis comprised: 1) rating the extent to which the article was accurate, complete, and so on on a seven-point Likert scale (ranging from "very low" to "very high") and 2) performing a count of the number of errors and omissions in the article. Next, the librarians met again, compared their quality assessments of the six "training" articles, and discussed differences. As a result, they articulated what constituted a low/medium/high error or omission and specific criteria for rating each of the quality dimensions along the seven-point scale. They defined low-rated errors or omissions as misspellings or mistakes in typography or grammar. They defined medium-rated errors or omissions as non-substantive but more significant than misspellings; in other words, these errors or omissions did not affect the understanding of a topic or entry (e.g., an incorrect number of siblings or whether or not the complete names were listed for parents). They defined high-rated errors or omissions as substantial errors or omissions that obscured the understanding of a subject. The criteria for accuracy and completeness was the following: 7 = perfect; 6 = very good (one or two low or non-substantive omissions or errors); 5 = good (more than two low or non-substantive errors or omissions such that the understanding was not obscured); 4 = pass (increased number of low or medium-grade errors or omissions that affected clarity of understanding); 3 = borderline (quantity and severity of errors or omissions affected understanding); 2 = fail (quantity and increasing severity of errors or omissions severely affected understanding); and 1 = disaster (errors or omissions make article completely inaccurate or incomplete). The librarians developed similar criteria for the other IQ constructs. We conducted this session over a week's time, and it required the librarians to devote two to three hours of independent work and roughly five hours of discussion (over multiple meetings).

Finally, the librarians actually began to assess the quality of the Wikipedia articles in the study set. When assessing the quality of Wikipedia articles, the librarians worked in a quiet office space where they had access to library resources and to the Internet. During this process, the librarians did not discuss with each other any of the Wikipedia entries, their own research of the topics, the sources, scoring, or any other aspect of the Wikipedia assessment. We also instructed the librarians not to review the current online version of the Wikipedia entry (to ensure they were all analyzing the exact same article version, or the history or discussion pages. The librarians worked in sessions of 1-2 hours per day to ensure that they stayed focused; on average, they spent 20 minutes per Wikipedia article (totaling 30-35 hours). We conducted the entire process over two summer months (when librarians had more time for research-based activities).

The data collected from the librarians included counts of errors and omissions for each of the Wikipedia articles in our set and ratings of their perceptions regarding the quality of the articles along the various IQ dimensions. Figure 1 illustrates the format of the collected data.



	Assessor #1						Assessor #2						Assessor #3					
	Counting		Rating Perceptions				Counting		Rating Perceptions				Counting		Rating Perceptions			
	Errors	Omissions	Accuracy	Completeness	Objectivity	Representation	Errors	Omissions	Accuracy	Completeness	Objectivity	Representation	Errors	Omissions	Accuracy	Completeness	Objectivity	Representation
Article #1	4	1	4/7	1/7	3/7	3/7	6	2	5/7	6/7	5/7	4/7	12	4	7/7	4/7	2/7	6/7
Article #2	11	7	5/7	6/7	5/7	4/7	1	7	7/7	4/7	2/7	6/7	5	8	4/7	1/7	3/7	3/7
…																		
…																		
…																		
Article #98	2	8	7/7	4/7	2/7	6/7	5	3	4/7	1/7	3/7	3/7	3	2	5/7	6/7	5/7	4/7

Figure B1. Format of the Data Collected

Once the librarians completed this extensive assessment procedure, we calculated their level of agreement over the entire set of Wikipedia articles for both the quality perception rating (on each of the quality dimensions) and the error and omission counts. For measuring inter-rater agreement, we used the intra-class correlation (ICC) and reliability of scale metrics. To illustrate the difference between these two metrics, consider the case of two assessors and four items, where assessor 1's rating vector was [1,2,3,4] and assessor 2's rating vector was [2,4,6,8]. In this case, scale reliability was very high (0.96) because the two vectors had a highly similar pattern, while intra-class agreement was mediocre (0.47) since absolute values differed.

To calculate the statistical significance of differences in inter-rater reliability, we followed Klein et al.'s (2001) and Wong's (2008) approach in which one calculates the standard deviation for each of the items (in our case, Wikipedia articles) that multiple assessors rated. We note that other methods for estimating whether inter-rater reliability scores are similar between pairs of observers who have evaluated the same set of items exist (cf. (Cicchetti & Heavens, 1981; Fleiss, Cohen, & Everitt, 1969). However, we could not use these methods in our case because they require the minimum number of observations to be 3*K2 (where K is the number of categories; in our case, we would have needed 147 articles (roughly 50% larger than our set of Wikipedia articles)). Once calculating the standard deviation in ratings for each article, we used this metric as an outcome variable and tested the significance of differences in means using the Mann-Whitney U test (2-sided). We repeated this calculation for all IQ dimensions.

Appendix C: Additional Details Regarding the Results for the Qualitative Analysis of the Application of Heuristic Principles in IQ Assessment

The manuscript body summarizes the results for the IQ dimensions of accuracy, completeness, objectivity, and representation. Below, we provide the detailed results for each one of these IQ dimensions.

Assessing Accuracy: Heuristic Principles

Search Directions

Participants began with an internal search in which they scanned over the text unsystematically, read only parts of the text, jumped around, and attended to cues as they encountered them. Given that searching in articles was typically unsystematic, different participants went over the text in different trajectories: some participants briefly scanned the text and others read some paragraphs carefully. When attending to cues in the focal article, a few specific cues (e.g., numerical data, references, dates) attracted most participants' attention. In addition, when assessing accuracy, most participants (22 out of 30) searched externally by consulting other Web resources and comparing their content to that of the focal Wikipedia article. We found that they used four different search strategies: keyword search (using a search engine) for the article's topic; a search with keywords related to a particular fact presented in the wiki article; a search for a specific target website, often an official source for that topic; and, less frequently, a search for a Wikipedia article on a related topic. In the case of the first two search directions (representing roughly two thirds of the participants), assessors commonly landed on the same two to three webpages, which resulted in high consistency in search direction. In contrast, when employing the latter two strategies (about a third of the cases), participants arrived at different webpages or ended up not using an external source (i.e., inconsistency in search direction). In addition to searching the focal Wikipedia article and external sources, in seven out of the 30 cases, participants relied on their own prior knowledge (or commented that they would have relied on it had they held domain-specific knowledge) or expectations. A few examples include comments such as: "I think that the article is quite extensive, because I am familiar [with this topic]", "I also relied on my previous knowledge and what seemed reasonable to me", and "I have no clue as to the accuracy of the article, since I have no prior knowledge on the topic".

Stopping Rules

Some of the participants stopped their search based on internal information alone, often in cases when they perceived specific cues (or anchors) as indicators for high accuracy. Common cues were the inclusion of references, dates, or numerical data. Examples of statements that participants made include: "the tables made me believe that I don't need to put much effort into assessing accuracy", "the numbers and references make the article seem reliable", and "the many numbers presented made a strong impression". Examples of less common cues used for terminating search include fluency of language and consistency of article's contents. When searching external Web sources, participants often terminated their search when encountering one or two reliable sources (considering, for example, websites of authoritative sources such as government agencies or articles written by domain experts as reliable). In cases where they found no reliable source, participants relied on sources perceived as less reliable (participants often scanned two or three such sources prior to stopping the search). When comparing the focal article to other sources, the participants typically checked for no more than two facts and a single fact found to be inaccurate resulted in their terminating the search. In a few cases, the participants terminated the search when they could not locate any relevant external sources.

Decision Rules

Participants' decisions regarding the accuracy of the article was highly influenced by the occurrence of several cues in the focal article: the presence and number of references, level of detail, numerical data, the presence of charts and tables, and high-quality grammar and style. Participants consistently interpreted these cues as contributing to the article`s accuracy. For example, one participant said:

> *I was influenced by the first impression that a thorough job was done in writing this article. It seemed to me that somebody really researched this topic, and the level of detail impressed me. All these gave me the impression that I can trust the accuracy of the [Wikipedia] article.*

In the cases when participants consulted external sources, the comparison to facts in external sources was the factor that received the highest weighting in the assessment of accuracy. The following cues found in external sources influenced participants' decision making: facts corroborated with other (preferably reliable) sources, not finding any incorrect fact, and external sources' being less detailed (when compared to focal Wikipedia article). Participants consistently interpreted these cues as contributing to the accuracy of the focal article. In the vast majority of cases, identifying a single corroborating fact led to the participants' perceiving the article as accurate; similarly, they weighted finding one or two inaccurate facts heavily and lead to their perceiving the article as inaccurate. One participant explained this heavy anchoring as follows: "If I found one incorrect fact, then there are probably many more [incorrect facts] in the article". Using a different decision rule, some participants considered a much larger set of cues and listed them all without indicating any differences in weighting. For example, when describing his decision making process, one of the participants listed many factors:

> *The systematic presentation of the company's milestones contributes to the information reliability…. The topic headings within the article look good, but within each topic, the content is not so impressive. Had the article been edited recently, it would have been much better; as it is, I think that it could be misleading for someone who has less knowledge about this company. …There are no outright lies, but the information is too outdated.*

Participants who relied on their own domain knowledge often judged the article's accuracy based on the extent to which it corroborated their prior knowledge of the topic. Similarly, participants' expectations played an important role in accuracy judgments—in particular, their disposition towards Wikipedia and the community-based knowledge co-production model. For example, one participant commented that: "I trust Wikipedia, and I was also impressed by this being the first result in Google" and indicated that he assessed the article as being highly accurate.

Assessing Completeness: Heuristic Principles

Search Directions

When analyzing the focal Wikipedia article, participants scanned over the text; they often read only some of the text and zoomed in on cues as they found them. Many participants started by scanning the list of topics and then proceeded to read relevant sections. In only about one third of the cases (11 out of 30) did participants search outside the focal article by comparing its contents to external sources (the participants consulted three distinct sources in those 11 cases), specifically seeking content that was excluded from the Wikipedia article. Participants also searched their personal knowledgebase, and, in close to half of the cases (14 out 30), participants referred to their expectations often based on personal domain knowledge regarding what should and should not be included the focal Wikipedia article. For example, while reading an article on a car manufacturer, one participant commented that "I would have liked to know what happened after the company went out of business". Another participant reading the article about a museum stated: "There is not enough information to learn about the museum, the artifacts it displays, what the building looks like, or how it is divided by floors. I would not know where to go or what to do there, based on reading the article.". Note that, although participants had consistently sought to identify facts or descriptions omitted from the Wikipedia article, the particular omissions they each identified differed considerably.

Stopping Rules

Most of the participants stopped their search based only on cues in the focal article. The most common cues that attracted participants' attention were the length and level of detail (which both served as indicators for completeness), and participants often made their assessment based on the table of contents. When encountering such cues, participants decided to stop their search while making comments such as: "Even merely based on text length, I would not think the article is complete. It looks as if many things are missing, but I don't really know what they are" or "The *completeness* level of the article is very high…[and] there are enough details". Less common cues for completeness were the inclusion of images and numerical data.

Decision Rules

Participants' decision on the completeness of the articles relied heavily on their length and level of detail and often based their impression on the table of contents: they interpreted lengthy text, many details, and a long list of topics (that included the topics the participants expected to find) as contributing to the article’s completeness. In the cases when participants consulted external sources, they assessed completeness was based on 1) whether they found content missing from the focal article in the external source and 2) the importance they assigned to that omission. Omitting a fundamental fact negatively influenced perceptions of completeness. For example, when assessing the museum’s Wikipedia page and consulting an external source, a participant commented: “I would expect the museum’s address, contact details, and additional information relevant to those who wish to visit the museum to be included in the Wikipedia article as well.”. In contrast, an insignificant omission did not affect completeness perceptions.

Assessing Objectivity: Heuristic Principles

Search Directions

In the vast majority of cases (25 out of 30), participants relied on internal search alone when evaluating the focal Wikipedia article for objectivity: they typically scanned the text top to bottom and paused on segments that caught their attention. Several kinds of text segments consistently attracted participants’ attention (e.g., the “pros and cons” section in controversial articles. Participants commonly looked for cues such as opinionated statements but usually found it difficult to discover cues for assessing the article’s objectivity. For example, one participant commented: “It is hard to find an indication for subjectivity.”. The few participants who compared the article to external sources (5 out of 30) typically relied on the same sources they had found when evaluating accuracy. In these cases, search direction was very focused and centered on a specific piece of information suspected as biased (e.g., a critical statement in the focal article). Overall, participants’ search directions were highly consistent because the majority restricted their search to the focal article, and the search in that article followed a similar pattern.

Stopping Rules

The participants who based their evaluation only on internal search scanned the article by starting with the first paragraph and anchoring on headings; they sometimes stopped at specific text segments that appeared potentially useful in helping to assess objectivity. One common cue was the level of perceived discourse in the article. As one participant said (while smiling): “I admit that I can be fooled by the use of highly formal language into believing that the text is objective—even if it isn’t.”. By and large, participants struggled to identify relevant cues: in 17 out of 30 cases, they identified no cues or relevant text segments, and they stopped the assessment when they had completed scanning the article. The few that extended their search beyond the focal article often terminated the search after identifying a single cue (e.g., finding a fact that either corroborated the focal article or contradicted it). Overall, the pattern of behavior when deciding when to stop the search was quite consistent.

Decision Rules

In cases in which participants found no explicit cue by which to assess objectivity (e.g., opinionated statements), they sometimes applied a default decision rule and perceived objectivity to be high while making comments such as: “The information was factual and not based on any opinion” or “the author did not state his own opinion”. At other times, participants could make a statement about the article’s objectivity, yet they were not able to distinguish any particular cue. When assessing a Wikipedia article about a particular car manufacturer, one participant commented:

> *It is hard for me to pinpoint why I feel that the article is so subjective. My guess is that the author bought a car and was not satisfied with it, or that he had something against the owners of the company.*

Some participants explicitly admitted that they were not sure what individual cues they should use to evaluate objectivity and relied mostly on the accumulated weight of particular kinds of information as a cue. For example, one participant explained his decision making process by commenting that: “since there were quite a few facts mentioned in this article, I saw no need to check them…. My hunch is that it’s objective, because I mostly noticed factual information.”.

In cases when participants did identify cues for objectivity, they found it difficult to interpret these cues, and, in some cases, a single cue that was interpreted as an indicator for bias by one participant was viewed by another participant as a marker of objectivity. For example, some viewed a section that criticized the entity described in the article as indicating a negative bias and by others as a sign for objectivity (balancing more positive statements). Similarly, some deemed facts that described a person were as factual and others deemed them too flattering. This difficulty in interpretation of cues for objectivity was also apparent when comparing the focal article to external sources. For example, some viewed an omission of a particular fact as a sign of incompleteness, while others perceived that same omission as deliberate (and, thus, as a sign of bias); they stated: “for all I know, the article may have deliberately omitted the information [about the particular fact]”. In another example, participants noticed that the content of the focal Wikipedia article that described a museum was quite similar to the museum’s website; some interpreted this as a sign of objectivity (a museum’s site was perceived as highly reliable), while others suspected that this may in fact contribute to bias (the museum’s site was promotional).

Assessing Representation: Heuristic Principles

Search Directions

When assessing the focal Wikipedia article for representation, participants scanned over the text and paid particular attention to visual aids (most notably, pictures, but also charts). In addition, several participants focused on the article structure (focusing on headers). Participants rarely consulted external sources (only four of 30 cases), and participants who searched external sources typically searched for visual aids and compared them to the visuals in the focal article.

Stopping Rules

Typically, participants identified only a few (two or three) cues before forming a perception of the article’s representation and terminating the search. The most salient cues were images and diagrams; other cues used for assessing representation included coherent organization by sections; intelligible text, language and articulation; and the use of summary statistics.

Decision Rules

The decision regarding the article’s representation relied heavily on the use of pictures and their number and contribution to the intelligibility of the article. When considering structure and style, participants weighted the article’s readability and comprehensibility. The use of these cues and their interpretation in terms of the assessment of representation was highly consistent across participants.

About the Authors

Ofer Arazy is a faculty member in the Department of Information Systems at the University of Haifa. He received his PhD from the University of British Columbia (UBC). His research interests—broadly speaking—are in the areas of online communities, knowledge management and computer supported cooperative work (CSCW). His research has been supported by various funding agencies and external sources. His work has appeared in the field's premier journals: *MIS Quarterly*, *Information Systems Research*, *Journal of MIS*, *Journal of the AIS*, and the *Journal of the Association of Information Science and Technology*, among others. His work has received accolades, including the 2010 AIS Best IS Publication of the Year Award.

Rick Kopak is Associate Director in the School of Library, Archival and Information Studies (iSchool) at the University of British Columbia. He received both his Masters and PhD in information studies at the University of Toronto. His research specialization is in Human Information Interaction with a particular interest in information design in digital reading environments. His research has been published in a variety of journals, including the *Journal of the Association of Information Science and Technology*, *International Journal of Information Management*, *First Monday*, and *ACM Computing Surveys*, and in array of conference proceedings including those for Information Interaction in Context, Human Computer Interaction for Information Retrieval, the ACM SIGIR Conference on Human Information Interaction and Retrieval, and the Association for Information Science and Technology.

Irit Hadar is a tenured faculty member at the Department of Information Systems, University of Haifa, and the Head of the Software Architecture Laboratory at the Caesarea Rothschild Institute for Interdisciplinary Applications of Computer Science. She received her PhD from the Technion—Israel Institute of Technology. Her main research area is cognitive aspects of requirements analysis, software architecture and design. Hadar has published over 90 papers in international journals and conferences (e.g., *CACM*, *JAIS, EJIS, IST*, *EMSE*, *REJ*, *JSS*, *RE*, OOPSLA, AMCIS, MCIS), has served as an organizer and PC member in various conferences and workshops, and has served as an editorial board member of the *ACM Transactions on Computing Education* (2011-2015).