On the Measurability of Information Quality

Ofer Arazy
Department of Accounting and Management Information Systems, Alberta School of Business, University of Alberta, Edmonton, AB T6G 2E8 Canada. E-mail: ofer.arazy@ualberta.ca

Rick Kopak
School of Library, Archival and Information Studies, University of British Columbia, Vancouver, British Columbia, V6T 1Z1 Canada. E-mail: rkopak@interchange.ubc.ca

The notion of information quality (IQ) has been investigated extensively in recent years. Much of this research has been aimed at conceptualizing IQ and its underlying dimensions (e.g., accuracy, completeness) and at developing instruments for measuring these quality dimensions. However, less attention has been given to the measurability of IQ. The objective of this study is to explore the extent to which a set of IQ dimensions—accuracy, completeness, objectivity, and representation—lend themselves to reliable measurement. By reliable measurement, we refer to the degree to which independent assessors are able to agree when rating objects on these various dimensions. Our study reveals that multiple assessors tend to agree more on certain dimensions (e.g., accuracy) while finding it more difficult to agree on others (e.g., completeness). We argue that differences in measurability stem from properties inherent to the quality dimension (i.e., the availability of heuristics that make the assessment more tangible) as well as on assessors' reliance on these cues. Implications for theory and practice are discussed.

Introduction

User assessment of the quality of Web-based information has received significant attention in the research-based literature over the past decade. Two major reasons for this attention are (a) the phenomenal growth in the number of information sources available on the Web and (b) the highly accessible nature of this information by a diverse set of consumers. With the diminution of traditional gatekeeping on the "information production" side (e.g., editorial and peer-review processes), more and more of the available content is obtained from sources with mixed, and sometimes dubious, provenance. A consequence of unreliable authority of sources and questionable quality of information is greater reliance on the ability of information consumers to make these quality judgments. Lankes (2008) described this as part of a larger trend toward "information self-sufficiency," where more and more of our everyday decision making is based on receiving information that is "disintermediated." The paradox resulting from this is that "end users are becoming more responsible for making information determinations, but because they have fewer physical cues to work with, they are becoming more dependent on the information provided to them by others." (Lankes, 2008, p. 104).

Information quality (IQ), as a concept, has been investigated extensively in prior information science research, where much of the discussion has been devoted to the underlying dimensions (or attributes) of IQ, such as accuracy, completeness, presentation, and objectivity (Hilligoss & Rieh, 2008; Lee, Strong, Kahn, & Wang, 2002; Liu, 2004; Rieh & Danielson, 2007; Wang & Strong, 1996). Largely, these investigations have focused on the salience of the various dimensions, studying whether one quality dimension better represents users' perceptions of IQ than does another dimension. These studies have shown that information consumers may perceive certain quality dimensions to be more important than are others, and for a variety of reasons, including domain expertise (Stanford, Tauber, Fogg, & Marable, 2002), gender (Flanagin & Metzger, 2003), or differences in information-seeking style (Rains & Karmikel, 2009).

The objective of this study is to investigate the consistency with which different users assess dimensions of IQ, and to compare consistency levels across the various dimensions. We argue that IQ dimensions, by their nature, differ in the extent to which they lend themselves to reliable measurement, such that when multiple assessors (as users or readers of the information) analyze a set of information objects, the level of agreement reached will vary depending on the quality

Received March 15, 2010; revised September 3, 2010; accepted September 16, 2010

 • Published online 5 November 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.21447

dimension they are asked about. We refer to this trait as the *measurability of an IQ construct*.[1]

By identifying those dimensions of quality that are more or less able to be judged consistently, we may learn more about what aspects of quality are easier (or more difficult) to assess than others. For example, some dimensions may be less context-sensitive (e.g., less task-dependent), relying more on extrinsic indicators that span across all tasks, and therefore make better "general-purpose" indicators of quality. This would have important implications for both research and practice. For empirical research on IQ, it is essential that we are aware of the limitations of existing measurement instruments. Generally speaking, reliability is the consistency of a measurement; it describes the degree to which an instrument measures the same way each time when it is used under the same conditions with the same subjects (Fleiss, 1986; Moskal, 2000). In our case, we are interested in the consistency between multiple assessors analyzing the same set of information objects. While prior studies have done a good job at ensuring convergent validity (i.e., the extent to which multiple items measuring the same construct are correlated), this is not sufficient for addressing the subjectivity that is associated with raters' interpretation of the information (Stemler, 2004) and judges' ability to recognize an object's quality. A valid measurement scale would produce consistent ratings between independent judges, or high interrater reliability, and this is of special importance when the nature of the phenomenon under investigation is difficult to observe. Oakleaf (2009), in her discussion of instruments for measuring information literacy, stated that prior studies have not paid attention to issues of interrater reliability. Similarly, prior research on IQ has paid little attention to the consistency in multiple users' quality perceptions and assessments of the various IQ dimensions. We argue that to draw any conclusion from studies on IQ, it is required that measurement instruments produce high interrater reliability.

The ability of assessors to agree on the quality of an information object has important implications for practice as well. Some Web services produce IQ metrics for their published content, and often these metrics are based on users' ratings (e.g., health-related Web sites). An understanding of which dimensions tend to produce higher agreement than others would have implications for a quality-assessment procedure (e.g., requiring more ratings for low-agreement dimensions) as well as for the presentation of information (e.g., indicating variance of scores, in addition to total quality score).

Our aim is to investigate the measurability of IQ constructs; that is, the extent to which existing scales of IQ dimensions lend themselves to consistent assessments by multiple judges. Our research question is whether there are some recognized dimensions of IQ that are inherently more reliable and that show less variation in terms of raters' agreement levels. To make sure that raters' agreement is an artifact of the specific domain of knowledge from which articles are pulled, we designed a comprehensive task where raters assessed a large and diverse set of information objects. We investigated the agreement in assessments for a sample of 270 undergraduate university students, where each student rated the quality (i.e., accuracy, completeness, objectivity, and representation) of 2 articles from a set of approximately 100, such that each article was rated by several students.

We chose Wikipedia articles as the content upon which users made their quality judgments for a number of reasons. First, and in keeping with its more general popularity, Wikipedia is a readily accessible information source and is available to all study participants. Since we were not concerned with investigating accessibility as an additional dimension of IQ in this study, Wikipedia was a natural choice. Wikipedia also is a source that is familiar and well-used by many, including those members of our sample. A recent Pew Internet and American Life Project study (Rainie & Tancer, 2007), for example, reported that over one third of the users of online resources polled consulted Wikipedia articles. Although there have been studies questioning the "quality" of the information contained in Wikipedia (Denning, Horning, Parnas, & Weinstein, 2005; Luyt, Aaron, Thian, & Hong, 2008; Wallace & Fleet, 2005), others have shown that overall, the quality of Wikipedia articles is quite good (Chesney, 2006; Stvilia, Twidale, Smith, & Gasser, 2008). Furthermore, Giles (2005) found that the quality of Wikipedia articles comes close to the quality of articles in Encyclopedia Britannica, and Fallis (2008) argued that "the epistemic consequences of people using Wikipedia as a source of information are likely to be quite good" (p. 1662). Lim (2009) reported that students in her study had generally "positive experiences" in using Wikipedia, and although they were aware of its limitations, they perceived that it was an adequate source in which to find "reasonably good" information (p. 2200).

The remainder of the article is organized as follows. We first review related research, and then describe our methods for estimating the interrater reliability of IQ measures. We follow this with a report on the results of our evaluation, and conclude with some reflections on the findings and a discussion of possible avenues for future research.

IQ and Its Assessment

In this section, related studies on IQ and its underlying dimensions are reviewed. Also discussed are users' perceptions and evaluations of IQ, and various issues relating to the reliability of users' assessments of IQ dimensions.

IQ Dimensions

As important and felicitous as it would be to have one, there is no generally agreed-upon definition of IQ (Michnik & Lo, 2009). As a concept, it is "elusive ... [and] of a

[1]Note that we are *not* referring to the reliability of the information itself or to the reliability of the information provider; this aspect has often been considered as one of the attributes of information quality. Instead, we are interested in the degree to which a construct lends itself to consistent measurement.

DOI: 10.1002/asi

transcendent quality (essence) synonymous with excellence" (Fink-Shamit & Bar-Ilan, 2008). Often, the definitions given are suggestive of a particular kind of utilitarian outcome. Taylor (1986), for example, saw quality as the value or worth the information has in relation to the purposes at hand. Alternatively, Hilligoss and Rieh (2008) emphasized the users' needs and his or her singular assessment, viewing IQ as the individuals' "subjective judgment of goodness and usefulness of information" (p. 1469). A more pragmatic approach might be to acknowledge both the "objective" and "subjective" views of IQ. Wang and Strong (1996, p. 6), for example, used as their definition of IQ "fitness for use," imbuing the definition with a sense of contingency, where quality depends on a judgment of value, or "fitness" of the information to a specific purpose or use. Eppler (2006) explicitly acknowledged the duality of the construct, and defined quality as the degree to which the information at hand either meets the requirements of the particular activity in which the user is engaged (the objective view) or the degree to which the information meets the expectations of the user (the subjective view). For the purposes of this investigation, we use the more general definition of quality—fitness for use—which encompasses both of these objective and subjective aspects.

The extensive literature on IQ recognizes that quality is a multidimensional construct, and has operationalized IQ through the use of specific attributes as indicators of its relative presence in information. Although there is much variation in the literature in the application of relevant nomenclature to describe the various "components" of IQ, quality indicators are often manifested at the primitive level as attributes. Furthermore, these attributes are often grouped into "quality dimensions"[2] comprising similar attributes. The groupings are made manifest in various ways ranging from more intuitive, manual classifications to the use of statistical methods such as factor analysis. Exemplary of the intuitively derived indicators are those of Taylor (1986), who identified five kinds of values (i.e., dimensions) that IQ may possess: accuracy, comprehensiveness, currency, reliability, and validity. As was typical of early efforts to identify the essential aspects of IQ, the dimensions were derived a priori. An alternative approach for defining IQ dimensions is through studies of user-based descriptions of quality (e.g., Wang & Strong, 1996). Whereas the intuitively derived classifications were obtained through reviewing prior literature, empirical studies engaged participants directly by soliciting attributes that were important in their individual perceptions of IQ. Wang and Strong's (1996) study, for example, surveyed 137 users, yielding 179 different quality attributes that eventually reduced to 20 dimensions, and then further reduced to four primary IQ "categories." More recently, several reviews have attempted to make IQ typologies more tractable and organized the various attributes and dimensions that have been used to operationalize IQ in the extant research literature. Lee et al. (2002) gathered IQ attributes from 15 (predominantly Management Information Systems) studies, differentiating between those studies employing attributes from academic and practitioner points of view. They adapted the categories proposed by Wang and Strong (1996) and reduced IQ attributes to four main categories. In a more recent review, Knight and Burn (2005) compared 12 earlier studies that used a variety of IQ attributes, reducing the number of attributes to 20 based on the frequency with which each attribute appeared across all of the studies examined.

In this preliminary investigation, we wished to explore the measurability of a restricted set of IQ dimensions rather than the gamut of IQ attributes and dimensions. The aim of our study was to investigate whether certain IQ dimensions are inherently more reliable than are others; that is, whether users have a higher level of agreement when asked to judge whether a particular piece of information is "accurate" as opposed to "complete." To aid in the selection of a restricted set of quality dimensions for our study, we employed Lee et al.'s (2002) categorization of derived dimensions. In their study (based on earlier work by Wang & Strong, 1996), four high-level categories that provided "comprehensive coverage of the multi-dimensional IQ construct" (Lee et al., 2002, p. 135) were empirically derived. Accordingly, *intrinsic* IQ represents dimensions that recognize that information may have innate correctness regardless of the context in which it is being used. For example, information may be more or less 'accurate' or 'unbiased' in its own right, or be characterized by the extent to which it conforms to true values or states in the world. *Contextual* IQ recognizes that perceived quality may vary according to the particular task at hand, and "must be relevant, timely, complete, and appropriate in terms of amount, so as to add value" to the purpose for which the information will be used (Lee et al., 2002, p. 135). *Representational* IQ addresses the degree to which the information being assessed is easy to understand and is presented in a clear manner that is concise and consistent. The fourth category, *Accessibility IQ*, references the ease with which the information sought is obtained, including the availability of the information and timeliness of its receipt. For our purposes, we chose to focus on three IQ categories from Lee et al.'s classification: "intrinsic," "contextual," and "representational" IQ. We disregarded "accessibility IQ," given that we considered only articles from Wikipedia, which is widely available on the Web and to which assessors all had easy access, such that there was little opportunity for variation in measurement across this category. In selecting specific IQ dimensions for our study from the intrinsic, contextual, and representational categories, we chose dimensions that reflected their frequency of occurrence within the studies examined in Lee et al.'s survey of the literature. The following IQ dimensions were chosen for our study: accuracy (intrinsic IQ), objectivity (intrinsic IQ), completeness (contextual IQ), and representation (representational IQ). We are not arguing that one information dimension is more appropriate or important than is another; rather, we selected a subset of dimensions that others have argued for as to their importance, and investigated the more

[2]In addition to dimensions, these groups of quality attributes also have been referred to as factors, categories, or criteria.

DOI: 10.1002/asi

specific concern of possible differences in the consistency with which users could assess quality within this subset of four IQ dimensions chosen.

Measurability of IQ Dimensions

To the best of our knowledge, no prior studies have explored the extent to which IQ dimensions lend themselves to consistent measurement, and to date, it is unclear whether multiple assessors would agree more on how "accurate" the information was, for example, than on how "objective" or "complete" it was. Note that the concept of interrater reliability at the heart of this study is fundamentally different from the notion of construct validity that is regularly investigated in empirical studies of IQ, although on the surface they bear some resemblance to one another. Construct validity refers to the degree to which inferences legitimately can be made from the operationalizations in a study to the theoretical constructs on which those operationalizations are based (Bagozzi, Yi, & Phillips, 1991). Two subtypes of construct validity are (a) convergent validity (the extent to which measures of the same construct are highly correlated) and (b) discriminant validity (the extent to which measures of distinct constructs are uncorrelated). To illustrate this idea, assume that Jack assesses the accuracy and completeness of a set of five articles, using two items to measure each construct. Using a Likert scale of 1 to 7, assume that Jack provides the ratings described on the left-hand side of Table 1. In this example, convergent validity is high since for any given article the scores of both measures of a construct (e.g., Acc1 and Acc2) are very similar. Discriminant validity also is high since the scores for "accuracy" and "objectivity" are clearly different. To illustrate how the notion of interrater reliability is different, assume now that Jill also rates the same set of articles, using the same measurement tool, and provides the rating described on the right-hand side of Table 1. In Jill's case, convergent validity is high, for the same reason as in Jack's example. However, the picture for interrater reliability is quite different. The interrater reliability of "accuracy" is low since Jack and Jill show no agreement (on any of the articles), but the reliability of "objectivity" is high since both assessors' ratings are consistent. Thus, construct validity is a prerequisite in the analysis of quality dimensions' interrater reliability, but it does not determine the interrater reliability scores. While construct validity is often analyzed in empirical studies of IQ (Flanagin & Metzger, 2003; Lee et al., 2002; Lim, 2009), it is not sufficient in cases when the nature of the phenomenon under investigation is difficult to observe and when there is a concern that independent judges would not agree in their assessments. Despite the importance the interrater reliability (LeBreton & Senter, 2008; Moskal, 2000), to date very little is known of the extent to which existing scales of IQ lend themselves to consistent assessments by independent judges.

Although prior studies have provided no explicit data indicating that a certain quality dimension is inherently more measurable or reliable than are others, extant literature does imply this. Differences in interrater reliability between various dimensions may stem from the availability of cues or the application of heuristics[3] based on certain structural aspects of the object that serve as more accessible indicators of a specific quality dimension (Hilligoss & Rieh, 2008). For example, the length of the article (i.e., the number of words) may serve as a cue for completeness while consistent headings could serve to indicate greater clarity in the representation of ideas within the article. When such heuristics are available to users, we expect that multiple assessors would reach higher levels of agreement. In reference to the categories proposed by Lee et al. (2002), it could be more likely that agreement would be high for the representational and contextual IQ categories, where cues are readily available, whereas interrater reliability would be lower for the intrinsic IQ category, where such cues are less apparent and where specialized knowledge may be required.

TABLE 1. Illustration of quality dimensions' multi-assessor reliability.

	Jack's Ratings				Jill's ratings			
	Accuracy		Objectivity		Accuracy		Objectivity	
Article	Acc1	Acc2	Obj1	Obj2	Acc1	Acc2	Obj1	Obj2
Article 1	7	7	1	1	2	2	1	1
Article 2	7	7	1	1	1	1	1	1
Article 3	4	4	7	7	1	1	7	7
Article 4	1	1	7	7	7	7	7	7
Article 5	1	1	4	4	7	7	4	4

In addition to studying the measurability of the four IQ dimensions—accuracy, objectivity, completeness, and representation—we also were interested in studying the extent to which a composite (or gestalt) IQ (CIQ) construct lends itself to consistent measurement across multiple assessors. We conjecture that it would be more difficult for multiple assessors to agree on such a high-level construct, as it is less straightforward to operationalize.

To summarize, to date little is known about the interrater reliability of IQ dimensions, and we can only conjecture which dimensions would result in higher agreement levels. Our investigation aims to fill this gap, and our research question concerns the differences in interrater reliability between four IQ dimensions: accuracy, objectivity, completeness, and representation. As an extension of this investigation, we also look at the reliability of an overall (CIQ) score.

Method

We employed a sample of 270 undergraduate student assessors that were recruited from a 3rd-year class at a North

[3]Hilligoss and Rieh (2008) identified three "levels of credibility judgments," of which heuristics is one. Similar to its use here, Hilligoss and Rieh defined heuristics as "general rules of thumb that are broadly applicable to a variety of situations" (p. 1473).

DOI: 10.1002/asi

TABLE 2. Items to measure information quality dimensions.

Construct	Code	Item description
Accuracy	Acc1	Information in the article is accurate.
	Acc2	Information in the article is correct.
Completeness	Comp1	The article includes all the necessary information.
	Comp2	The article is complete.
Objectivity	Obj1	The article is objective.
	Obj2	The article provides an impartial view of the topic.
Representation	Rep1	The article is clear and easy to understand.
	Rep2	The article is presented consistently.
	Rep3	The article is formatted concisely.
Composite Information Quality	CIQ1	The article is of high quality.
	CIQ2	The article provides a good description of the topic.

American university's business school. The majority of students were in their early 20s, with a near even male–female distribution. To measure assessors' interrater reliability, we asked the participants to independently assess the quality (along the various dimensions discussed earlier) of a series of information objects, and then compared their assessments. Assessors rated statements pertaining to the various quality dimensions on a Likert scale of 1 (*Strongly Disagree*) to 7 (*Strongly Agree*), using the set of items described in Table 2.

To ensure that quality assessments were not biased by variations in levels of domain knowledge, the set of information objects assembled included a broad representation of articles from the English-language version of Wikipedia (Nov, 2007). We employed a stratified sampling approach to represent the range of Wikipedia topics. We built on Wikipedia's top-level classification[4] (Kittur, Suh, & Chi, 2009) and further constructed a smaller set of six mutually exclusive and collectively exhaustive classes: (a) culture, art, and religion,[5] (b) math, science, and technology,[6] (c) geography and places, (d) people and self, (e) society,[7] and (f) history and events. We randomly selected 17 articles from each of these topical classes, with some restrictions. Since Wikipedia articles are often created as "stubs"—placeholders for further development—with little content, we included only articles that have passed the stub inception phase, and we set a lower limit of 200 words on article length. In addition, we were concerned that the effort required for assessing the quality for very long articles may bias assessors' ratings; therefore, we set an upper limit of 3,500 words on article length, thus excluding lengthy outliers. Examples of Wikipedia articles included in this procedure are: Bricriu (troublemaker and poet in Irish mythology), Dhol (a drum used in India), and Jacobi identity (in mathematics).

We employed a multistep research design to ensure that we obtained several assessments for each article in the set, yet constrain the amount of work involved in the task. This process is illustrated in Figure 1. As part of a class assignment, students were randomly assigned to Wikipedia articles and asked to assess the articles' quality. Since the thorough assessment of each Wikipedia article required significant effort, we randomly assigned each student to only two articles from the set, resulting in five to six different assessments for each article (see Step A in Figure 1). The students were instructed to study the contents of the Wikipedia articles assigned to them, compare each to alternative sources, and to consider the authority of these sources. They were then asked to judge each article's quality along the dimensions of accuracy, completeness, objectivity, representation, and CIQ, and represent their level of agreement with corresponding statements of quality on a Likert scale of 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). This process ensured that interrater agreement measures for the various quality dimensions were comparable since for every student–article pair, there existed an assessment along each of the dimensions. Students prepared a detailed report summarizing their analysis, in addition to the quality ratings, and were marked based on the depth of the report and the type of resources they employed. Of the 300 students in the class, 270 gave permission for their assignments to be used, and only these were included in the study. As a result, articles in our set had between one and six ratings each. We dropped articles with very few (<3) ratings, arriving at a set of 98 articles, each having three to six assessments (see Step B in Figure 1). To ensure that interrater reliability calculations were consistent across the entire set of Wikipedia articles (i.e., based on an equal number of assessments), we employed a K-fold cross-validation procedure (Kohavi, 1995). We produced 10 sets, each containing three assessments on all Wikipedia articles. If an article was rated by 4, 5, or 6 students, then we would randomly select three student ratings for that article (see Step C in Figure 1). Interrater agreement was calculated independently for each of the 10 sets, and we employed the average of these 10 calculations in our analysis.

In calculating interrater agreement, we first validated that the questionnaire items described in Table 2 indeed represented the IQ dimensions of interest and that construct validity was good (discussed earlier). Then, for each construct, we calculated an average score (e.g., the *Accuracy* score is based on the average of items *Acc1* and *Acc2*), and we employed these average scores in estimating interrater reliability. Recall that interrater reliability (also referred to as "interrater agreement") is the degree of agreement or consistency among raters. It gives a score of how much homogeneity is in the ratings given by judges. Interrater reliability is often

[4]For a list of Wikipedia top-level categories, please refer to http://en.wikipedia.org/wiki/List_of_overviews

[5]Our "culture, arts, and religion" class corresponds to the following Wikipedia categories: "Culture and the arts," "Religion and belief systems," and "Philosophy and thinking."

[6]Our "math, science, and technology" class corresponds to the following Wikipedia categories: "Mathematics and logic," "Natural and physical sciences," and "Technology and applied sciences."

[7]Our "society" class corresponds to the following Wikipedia categories: "Society and social sciences" and "Health and fitness."

DOI: 10.1002/asi

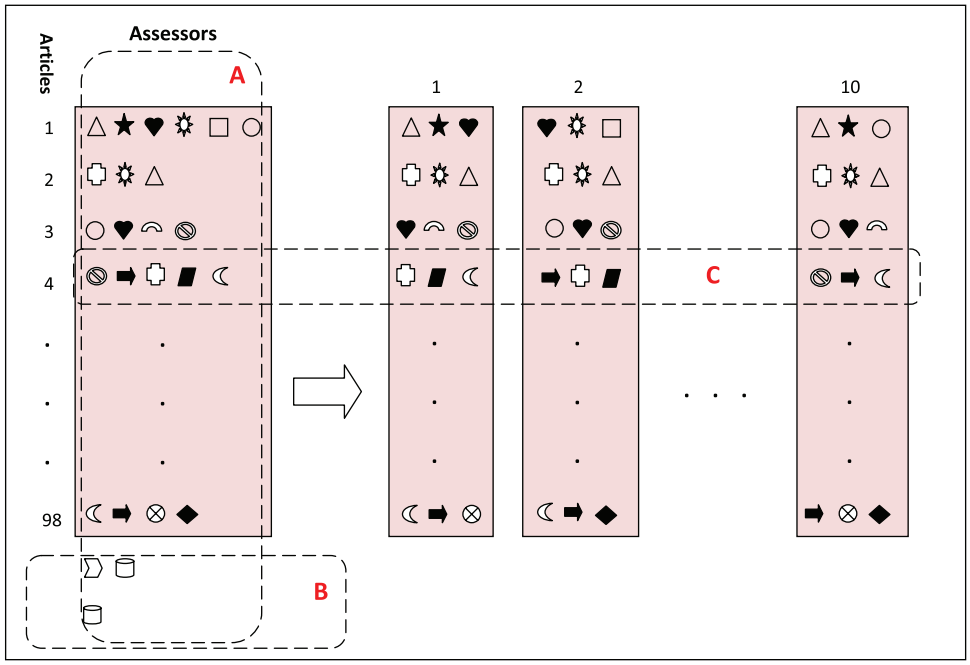

FIG. 1. Sampling procedure. Assessors analyzed Wikipedia articles from a predefined set. Each of the distinct shapes represents a unique assessor. A large number of assessors were each assigned to two articles (Step A), and articles with less than three assessments were removed from the procedure, leaving 98 articles (Step B). Next, we produced 10 sets of 3 raters' assessments for each of the remaining articles; in cases where an article was assessed more than three times, for each of the 10 sets we randomly selected three assessments (Step C). Interrater agreement was calculated independently for each of the 10 sets, and we used the average in our analysis.

employed for testing the tools given to human judges, for example, by determining if a particular scale is appropriate for measuring a particular variable (e.g., Yau, Etzkorn, & Virani, 2008). There are a number of statistics which can be used to determine interrater reliability. The most widely used measures in the behavioral sciences are the kappa measures (Kar & Yang, 2006), which has been recently employed in the field of information science (Oakleaf, 2009). Cohen's κ (Cohen, 1960), and its extension to more than two raters—Fleiss' κ (Fleiss & Cohen, 1973)—take into account the amount of agreement that could be expected to occur through chance.

In this study, we used the intraclass correlation (ICC) statistic (Haggard, 1958; Landis & Koch, 1977), which is directly analogous to Fleiss' κ (Fleiss, 1981; Fleiss & Cohen, 1973). Specifically, we used the intraclass agreement metric, range [−1,1], which emphasizes actual agreement on rating values. Furthermore, to detect cases where assessments differ, yet are 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 α indicator (which is commonly employed to estimate reliability of instruments). To illustrate the difference between these two metrics, consider the case of 2 assessors and four items, where Assessor 1's rating vector is [1,2,3,4] and Assessor 2's rating vector is [2,4,6,8]. In this case, scale reliability is very high (0.96) since the two vectors have a highly similar pattern while intraclass agreement is mediocre (0.47) since absolute values differ.

Our method for calculating the statistical significance of differences in interrater reliability follows the approach employed by Klein, Conn, Smith, and Sorra (2001) and Wong (2008), where the *SD* is calculated for each of the items (in our case, Wikipedia article) that was rated by multiple assessors. We independently repeated this calculation for each of the IQ dimensions. We then used the assessments' *SD* as an outcome variable and tested the significance of differences in means using a paired-sample test (two-sided).

Results

Instrument Validation

To validate our measures, we conducted a principle component analysis (PCA) with varimax rotation using SPSS. This produced a four-factor solution, corresponding to the four quality dimensions we have investigated: Accuracy, Completeness, Objectivity, and Representation. All items for these dimensions were found to have higher than 0.7 factor loadings and less than 0.3 cross-loading in the PCA. Items for CIQ did not produce a distinct factor but rather loaded on the other factors. Specifically, the CIQ items loaded on the factors corresponding to Accuracy (loadings of 0.34 and 0.56) and Completeness (loadings of 0.72 and 0.58), and to a lesser extent on the factor corresponding to Representation (loadings of 0.26), suggesting that CIQ is indeed a higher level construct that encompasses the dimensions of Accuracy, Completeness, and Representation. Interestingly, our

DOI: 10.1002/asi

TABLE 3. Item means, *SD*s, and factor loadings.

Construct	Item	*M*	*SD*	Factor1	Factor2	Factor3	Factor4
Accuracy	Acc1	5.20	1.31	0.93			
	Acc2	5.16	1.27	0.89			
Completeness	Comp1	3.72	1.72		0.90		
	Comp2	3.65	1.69		0.89		
Objectivity	Obj1	5.20	1.51			0.85	
	Obj2	5.21	1.51			0.89	
Representation	Rep1	5.38	1.59				0.83
	Rep2	5.40	1.31				0.80
	Rep3	5.45	1.38				0.84
CIQ	CIQ1	4.78	1.47	0.56	0.58		
	CIQ2	5.01	1.51	0.34	0.72		

Note. Factor loadings below 0.30 were suppressed.

TABLE 4. Means, *SD*s, reliability, intercorrelations, and average variance extracted.

Construct	*M*	*SD*	α	1	2	3	4	5
1. Accuracy	5.18	1.23	0.91	0.96				
2. Completeness	3.68	1.62	0.90	0.36**	0.89			
3. Objectivity	5.20	1.35	0.74	0.26**	0.24**	0.95		
4. Representation	5.41	1.22	0.81	0.31**	0.38**	0.27**	0.86	
5. CIQ	4.89	1.36	0.80	0.58**	0.68**	0.27**	0.48**	0.91

Note. The diagonals are the square root of the average variance extracted (AVE) for each of the factors.
**Significant at the 0.01 level (two-tailed).

data suggest that our assessors did not perceive Objectivity to be a dimension of CIQ but rather an independent construct (i.e., CIQ's loadings on Factor 4 are low). The four-factor solution explained 79% of the total variance. Table 3 presents the mean, *SD*, and factor loading of each measurement item.

To further assess construct validity, we created variables corresponding to each of the constructs by averaging the corresponding items. The average variance extracted (AVE; Fornell & Larcker, 1981) for the constructs ranged between 0.73 and 0.91, well above the 0.50 threshold. The square root of AVE for each construct was substantially higher than the correlation of the construct with other factors, demonstrating discriminant and convergent validity (Straub, Boudreau, & Geffen, 2004). All constructs have Cronbach's αs that satisfy the generally agreed-upon lower limit of 0.70 for confirmatory research (Straub et al., 2004), indicating that all measures are reliable. Table 4 presents the intercorrelations among the variables and their AVEs.

Interrater Reliability Results

In analyzing interrater agreement of IQ measures, for each measure we used the average rating of the corresponding items, as illustrated in Table 5. Generally speaking, we found that interrater agreement levels were low. Landis and Koch (1977) provided a scale for interpreting κ interrater value. A similar interpretation of ICC values was made by Fleiss (1981; Fleiss & Cohen, 1973). Their scale suggested that values below 0.20 represent "poor agreement," 0.21 to 0.40 "fair agreement," 0.41 to 0.60 "moderate agreement," and 0.61 to 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 could be attributed to the large number (i.e., 7) of categories we employed. Internal consistency—as measured through scale reliability—was higher than ICC, with values in the range of 0.18 to 0.36 (see Table 5); however, these values still represent only moderate agreement and also were likely influenced by the relatively large number of categories we employed.

When analyzing the differences in interrater reliability between the various quality dimensions, note that in terms of ICC, the highest agreement level was attained for Completeness, followed by Representation, with the lowest scores for Accuracy and Objectivity. The scale reliability results are consistent with the ICC results (see Table 5). The statistical significance of differences in interrater agreement was assessed based on the standard deviations in raters' assessments (using a paired-sample, two-tailed *t* test). The differences between all constructs' agreement levels, except for Accuracy–Objectivity and Accuracy–Representation, were statistically significant (at $p < 0.01$ or better). The analysis of interrater reliability for the CIQ construct revealed some interesting findings. In terms of both ICC (0.17) and scale reliability (0.38) , CIQ yielded a higher agreement score than all other dimensions (Differences from Completeness and

DOI: 10.1002/asi

TABLE 5. Interrater agreement results for the various constructs.

Users	Intraclass Agreement	Variance			Common Interitem Correlation	Reliability of Scale
		Common	True	Error		
Accuracy	**0.06**	1.80	0.17	1.68	0.06	**0.18**
Completeness	**0.16**	2.83	0.44	2.39	0.16	**0.36**
Objectivity	**0.10**	2.04	0.20	1.85	0.10	**0.26**
Representation	**0.14**	1.83	0.26	1.57	0.14	**0.33**

Note. Metrics we focus on are written in bold. Values of reliability of scale are unbiased.

Representation were statistically significant at $p < 0.001$ and $p < 0.05$, respectively; CIQ's differences from Accuracy and Objectivity did not reach significance levels.)

Discussion

In this article, we investigated the measurability of IQ dimensions and tested the extent to which specific groups of users of online information agree in their assessment of the resource's quality. To ensure that the agreement is not affected by the topic of the information objects, we designed an extensive information analysis task, consisting of the evaluation and rating of a large and diverse set of Wikipedia articles. We studied undergraduate university students' perceptions' of close to 100 articles in terms of four IQ dimensions: accuracy, completeness, objectivity, and representation. To explore the consistency in measurement of IQ dimensions, we performed a study of interrater reliability. For each Wikipedia article in our set, we had several quality assessments along the various dimensions, and we measured assessors' agreement. So, if the variance for objectivity is greater than that for completeness, we can say that objectivity as an indicator is more difficult to measure than is completeness, and hence the measure is less reliable. Our findings indicate that there are substantial differences between interrater reliability scores for the different quality dimensions, such that there is less consistency in the ratings of some indicators when compared to others.

The most striking finding from our study is that interrater reliability levels were very low—across all quality dimensions. Intraclass agreement levels did not exceed 0.17, indicating poor interrater reliability. For a comparison, a recent study of information literacy measurement tools, reported that in most cases the interrater reliability levels were in the 0.2 to 0.6 range (Oakleaf, 2009). The low agreement in our study could be attributed in part to the large number of categories (i.e., 7) in our scale; however, the low–moderate scale reliability scores suggest that the inconsistencies are more fundamental. Commonly, studies of IQ verify construct validity; however, most studies rely on the ratings of a single assessor. In those cases where more than one assessor rates each item, often the average assessors' ratings are used as a measure of quality without first testing interrater reliability. The interrater reliability values observed in our study were below the acceptable threshold and would not permit using assessors' average ratings.

The primary contribution of this study is in revealing the differences in interrater reliability between the various dimensions, demonstrating that some quality dimensions yield high agreement levels (completeness and representation) while others yield low agreement (accuracy and objectivity). We believe that these findings stem from the measurability or ease of assessment; that is, from the fact that for assessing certain dimensions there exist quick heuristics while for others there are none, or at least they are more difficult to understand or identify (cf. Hilligoss & Rieh, 2008). Specifically, an easy heuristic for completeness would be the quantity of content (e.g., the length of the Wikipedia article, or the presence of footnotes and bibliography). Similarly, representation may be easier to assess, and could be estimated based on consistency in structure and page design (Flanagin & Metzger, 2007). In contrast, no such straightforward heuristics are available for accuracy and objectivity, as their assessment requires a detailed reading of the content and a certain degree of domain expertise. In reference to Lee et al.'s (2002) conceptualization of IQ, intrinsic IQ (accuracy and objectivity) measures do not lend themselves to consistent rating, probably due to the lack of external cues, while contextual IQ (completeness) and representational IQ (representation) do yield relatively consistent ratings because—we believe—of the availability of heuristics.

Another important factor is the effect of domain expertise: Assessment of some quality dimensions requires less domain expertise than do others, resulting in more consistent ratings. Namely, the assessment of accuracy requires knowledge of relevant facts (or alternatively, a comparison of the information to external resources) while rating representation does not require such expertise. For example, assessors rated completeness based on the existence of specific sections of the article that they deemed important (Such an analysis does not require domain expertise.) We believe that when no heuristics are available and the ratings require detailed analysis, differences in skills and background become more salient, and results, we believe, in the inconsistent assessments and lower interrater reliability.

A secondary contribution of our study concerns the composite construct of IQ. While various studies refer to IQ as a composite construct that includes various dimensions, we are not aware of any studies that empirically analyzed the relations between this composite construct and its underlying dimensions. Our findings suggest that IQ incorporates

DOI: 10.1002/asi

various dimensions, including those of accuracy, completeness, and representation. This finding is in line with the extant literature. Interestingly, our factor analysis shows that the CIQ construct does not incorporate the dimension of objectivity, suggesting that objectivity may be perceived as a distinct dimension or at least one that can be evaluated separately from overall assessments of quality. Note that the agreement levels for CIQ were higher than were the levels recorded for other dimensions (although not all differences were statistically significant). Thus, it seems that the assessors could accommodate themselves to this composite concept and find heuristics to help estimate it. We suspect that the concept of IQ was intuitive for student assessors, who likely relied on the heuristics they employed for assessing completeness and representation in their assessment of CIQ.

Finally, our novel research methodology is in itself a contribution. The design for the study provides a framework for assessing interrater reliability for comprehensive evaluation tasks (i.e., tasks that require the assessment of many articles), where each article is assessed by a different assessor. Our design allocated several assessors to each item, where each assessor analyzed only a few items, and each item was assessed by multiple users. To handle the concern regarding the different number of assessors per item, we employed the K-fold approach, producing several "folds," each with the same number of assessors over all items. We expect that our methodology could generalize to the study of interrater reliability with other constructs.

Conclusion

The notion of IQ is of primary concern to information science scholars, and it has attracted significant attention in recent years. Various conceptualizations of IQ have been proposed, and most frameworks concur that IQ is a high-level construct that incorporates several dimensions (i.e., other constructs) such as accuracy and completeness. However, less attention has been given to the "measurability" (i.e., the ability to consistently measure) of IQ. Empirical studies of IQ often employ a survey to assess readers' perceptions of a resource's quality. Thus, the measurement of these quality constructs has been based on people's perceptions or estimates of appropriateness. Often, studies of IQ assume that people's abilities to perceive various dimensions are similar for all quality dimensions, and overlook issues of reliability of measurement. Findings from this study demonstrate the difficulty of reaching a consensus on IQ assessment, and reveal some important differences in agreement levels between these dimensions.

Implications for Research and Practice

Our findings have implications for both research and practice. The primary implication for information science scholars is the need for care in assessing IQ constructs. Using multiple items for constructs and ensuring the correlations between these items (i.e., ensuring construct validity) may not be sufficient, as there are likely to be inconsistencies between assessors in their perceptions of an object's quality. Since some quality dimensions are more difficult for assessors to agree on than are others (i.e., accuracy, objectivity), it is recommended that future studies of IQ give extra attention to the measurement of these constructs. Possibly, assessors could be given more training and allowed more time in making judgments on these dimensions, survey questions could be more specific, and more assessors could be employed, including measurement of individual domain knowledge and task-expertise levels that affect an assessor's ability to make judgments on the various IQ dimensions.

For information users, we would recommend care in judging quality, and in accepting others' quality ratings, as quality is such a highly subjective concept. Information users should realize that they (often unconsciously) employ heuristics in assessing quality, and that these heuristics are limited in estimating quality dimensions such as accuracy and objectivity. While knowledge of available heuristics for various IQ dimensions may be very useful, users should be aware of the limitations of these heuristics and that they may provide only a partial and somewhat limited indication of the overall quality of the object. Knowledge about these limitations also is important for information literacy education, where a greater focus might be placed on assessment techniques for those IQ dimensions less amenable to heuristic representation.

Another practical recommendation is aimed at Web services that produce IQ metrics for their published content. These metrics often are based on users' ratings. For example, many health-related Web sites have tools for estimating the quality of Web pages, and use symbols such as "award" or "seal" to indicate high-quality pages. These tools rarely report on the interrater reliability of the ratings (Gagliardi & Jadad, 2002). The low agreement levels recorded in our study suggest that ratings from a relatively large number of users are required for producing a quality score. Moreover, the differences in agreement for the various dimensions imply that users should be allowed to rate an article along various dimensions, and that more care should be placed (e.g., provide more guidance, require more raters) on the dimensions that are difficult to assess: accuracy and objectivity. One example in this direction is provided by the Public Library of Science (PLoS) journals. In PLoS, readers can rate an article according to insight, reliability, and style as well as a check box where you can indicate if you have any competing interests with the article (i.e., objectivity). Our suggestion to PLoS would be to allow readers to rate the articles on additional dimensions such as accuracy and completeness, and to consider the variance in responses when producing an aggregate quality score. Included with this might be a declaration (i.e., a self-assessment) of the rater's own level of expertise in the topical area addressed in the article being rated. Users of such services should be careful to accept quality scores without knowledge of what quality dimensions the score represents and the number of ratings used to generate it.

DOI: 10.1002/asi

Limitations and Future Research

Our study provides only preliminary findings regarding the measurability of IQ measures, and further research is warranted. First, there may be some concerns regarding potential biases in our sample (e.g., assessors' expertise or background); however, the design of our study addresses many of these concerns. We ensured that the set of Wikipedia articles we used in this study covered the spectrum of topics within Wikipedia; with such a wide range of topics, it is unlikely that one assessor had substantially more domain expertise than the others. In addition, our research questions concerned the differences in agreement between quality dimension (and not specific topics); thus, even if one assessor had superior domain knowledge of specific articles, this would manifest itself in his or her rating across *all* quality dimensions (e.g., accuracy, completeness, etc.). Therefore, differences in domain expertise are not expected to affect the comparative agreement levels between quality dimensions. 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 looks at the correlation between the assessors' ratings rather than the agreement of the actual values; having one assessor consistently apply more/less strict standards, thus, would not affect this metric. In the future, we hope to repeat our study with a larger sample size, directly measuring and controlling for exogenous factors such as assessors' cognitive or demographic traits (e.g., age, computer self-efficacy, information literacy, domain knowledge).

A second limitation of our study is that it investigated one information resource, Wikipedia, and it is possible that our findings were affected by assessors' biases or predispositions toward this resource. In the future, we plan to repeat this study on alternative information sources. Finally, the set of quality constructs employed in our study provides only a partial representation of this multidimensional construct, and we propose that future studies expand our investigation to additional quality dimensions (e.g., timeliness, understandability).

We conclude that IQ is an elusive construct that is hard to measure, and users' quality estimates are subjective, therefore making it difficult for multiple assessors to reach an agreement on a resource's quality. However, our study provides novel insights regarding the reliability of various IQ constructs that can be utilized in mitigating the less useful outcomes of this subjectivity. Still, additional research is required to validate our findings in alternative settings, expanding the scope of investigation, and to explore the role of additional factors that affect variations in agreement levels. We hope that our study will open the door for further research in this area.

Acknowledgments

We thank Kevin O'Kell for his work on this project. This research was funded in part by the Canadian Social Sciences and Humanities Research Council.

References

Bagozzi, R.P., Yi, Y., & Phillips, L.W. (1991). Assessing construct validity in organizational research. Administrative Science Quarterly, 36(3), 421–458.

Chesney, T. (2006). An empirical examination of Wikipedia credibility. First Monday, 11(11). Retrieved October 28, 2010, from http://www.firstmonday.org/issues/issue11_11/chesney/

Cohen, J. (1960). A coefficient for agreement for nominal scales. Education and Psychological Measurement, 20, 37–46.

Denning, P., Horning, J., Parnas, D., & Weinstein, L. (2005). Wikipedia risks. Communications of the ACM, 48(12), 152–152.

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

Fallis, D. (2008). Towards 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 behaviour. Information Research, 13(4), Paper No. 357.

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, 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). Statistical methods for rates and proportions (2nd ed.). New York: Wiley.

Fleiss, J.L. (1986). Reliability of measurement: The design and analysis of clinical experiments. New York: 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, 613–619.

Fornell, C., & Larker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18(1), 39–50.

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. British Medical Journal, 324, 569–573.

Giles, J. (2005). Internet encyclopedias go head to head. Nature, 438(15), 900–901.

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

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.

Kar, W.L., & Yang, C.C. (2006). Conceptual analysis of parallel corpus collected from the Web. Journal of the American Society for Information Science and Technology, 57(5), 632–644.

Kittur, A., Suh, B., & Chi, E.H. (2009). What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. In Proceedings of the 27th Annual Conference on Human Factors in Computing Systems (pp. 1509–1512). New York: ACM Press.

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, 3–16.

Knight, S., & Burn, J. (2005). Developing a framework for assessing information quality on the world wide web. Informing Science, 8, 159–172.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2(12), 1137–1143. San Mateo, CA: Kaufmann.

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

DOI: 10.1002/asi

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: MIT Press.

LeBreton, J., & Senter, J. (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.

Liu, Z.M. (2004). Perceptions of credibility of scholarly information on the web. Information Processing & Management, 40(6), 1027–1038.

Luyt, B., Aaron, T., 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.

Michnik, J., & Lo, M. (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 and Evaluation, 7(3).

Nov, O. (2007). What motivates Wikipedians? Communications of the ACM, 50(11), 60–64.

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.

Rainie, L., & Tancer, B. (2007). Wikipedia users. Pew Internet and American Life Project. Retrieved January 13, 2010, from http://www.pewinternet.org/Reports/2007/Wikipedia-users.aspx

Rains, S.A., & Karmikel, C.A. (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, 544–553.

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

Sim, J., & Wright, C.C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Physical Therapy, 85(3), 206–282.

Stanford, J., Tauber, E.R., Fogg, B.J., & Marable, L. (2002). Experts vs. online consumers: A comparative credibility study of health and finance web sites. Retrieved January 13, 2010, from http://www.consumerwebwatch.org/dynamic/web-credibility-reports-experts-vs-online-abstract.cfm

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment Research and Evaluation, 9(4).

Straub, D.W., Boudreau, M.-C., & Gefen, D. (2004). Validation guidelines for IS positivist research. Communications of the Association for Information Systems, 13, 380–427.

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

Taylor, R.S. (1986). Value-added processes in information systems. Norwood, NJ: Ablex.

Wallace, D.P., & Fleet, C.V. (2005). The democratization of information? Wikipedia as a reference resource. Reference & User Services Quarterly, 45, 100–103.

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.

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, 591–614.

Yao, H., Etzkorn, L., & Virani, S. (2008). Automated classification and retrieval of reusable software components. Journal of the American Society for Information Science and Technology, 59(4), 613–627.

DOI: 10.1002/asi