Published:

Arazy O. and Kopak R., 2011, On the Measurability of Information Quality, *Journal of the American Society for Information Science and Technology (JASIST)*, 62(1), pp. 89-99.

On the Measurability of Information Quality

Abstract

The notion of information quality has been investigated extensively in recent years. Much of this research was aimed at conceptualizing information quality 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 information quality. The objective of this study is to explore the extent to which a set of information quality 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 of these 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.

Keywords: information quality, measurability, reliability, accuracy, completeness, objectivity, representation, inter-rater reliability.

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) describes 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 is devoted to the underlying dimensions (or attributes) of information quality, e.g. accuracy, completeness, presentation, objectivity (Hilligoss & Rieh, 2008; Rieh & Danielson, 2007; Liu, 2004; Lee, et al., 2002; 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 information quality than another dimension. These studies show that information consumers may perceive certain quality dimensions to be more important than others, and for a variety of reasons, including domain expertise (Stanford et al., 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 information quality, and to compare consistency levels across the various dimensions. We argue that information quality 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 dimension they are asked about. We refer to this trait as the *measurability of an information quality 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 information quality, 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 do 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 rater's 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 inter-rater 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, states that prior studies have not paid attention to issues of inter-rater reliability. Similarly, prior research on information quality has paid little attention to the consistency in multiple users' quality perceptions and assessments of the various information quality dimensions. We argue that in order to draw any conclusion from studies on information quality, it is required that measurement instruments produce high inter-rater 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 information quality 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 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, and addition to total quality score).

Our aim is to investigate the measurability of information quality constructs, i.e. the extent to which existing scales of information quality dimensions lend themselves to consistent assessments by multiple judges. Our research question is whether there are some recognized dimensions of information quality that are inherently more reliable and that show less variation in terms of raters' agreement levels. In order 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, hence available to all of the 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 is also 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 a 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 et al., 2005; Wallace & Fleet, 2005; Luyt et al., 2008), others have shown that, overall, the quality of Wikipedia articles is quite good (Chesney, 2006; Stvilia et al., 2008). Furthermore, Giles (2005) found that the quality of Wikipedia articles comes close to the quality of articles in Encyclopedia Britannica, and Fallis (2008) argues 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 paper is organized as follows. We first review related research, and then proceed to describe our methods for estimating the inter-rater reliability of information quality 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.

Information Quality and Its Assessment

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

Information Quality Dimensions

As important, and felicitous as it would be to have one, there is no generally agreed upon definition of information quality (Michnik & Lo, 2009). As a concept, it is "elusive . . . [and] of a 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, sees quality as the value or worth the information has in relation to the purposes at hand. Alternatively, Hilligoss & Rieh (2008) emphasize the users' needs and his or her singular assessment, viewing information quality 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 information quality. Wang and Strong (1996), for example, use as their definition of information quality, "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 (2003) explicitly acknowledges the duality of the construct and defines quality as the degree to which the information at hand either meets the requirements of the particular activity the user is engaged in (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 information quality recognizes that quality is a multi-dimensional construct, and has operationalized information quality 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, through 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 value (i.e. dimensions) that information quality may possess: accuracy, comprehensiveness, currency, reliability, and validity. As typical of early efforts to identify the essential aspects of information quality, the dimensions were derived *a priori*. An alternative approach for defining information quality dimensions is through studies of user-based descriptions of quality (e.g. Wang and 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 information quality. The Wang and Strong (1996) study, for example, surveyed 137 users, yielding 179 different quality attributes, eventually reduced to twenty dimensions, and then further reduced to four primary information quality ‘categories’. More recently, several reviews attempted to make information quality typologies more tractable, and organized the various attributes and dimensions that have been used to operationalize information quality in the extant research literature. Lee et al. (2002) gathered information quality attributes from fifteen (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 information quality attributes to four main categories. In a more recent review, Knight and Burn (2005) compared twelve earlier studies that used a variety of information quality attributes, reducing the number of attributes to twenty, 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 information quality dimensions, rather than to try and cover the gamut of information quality attributes and dimensions. The aim of our study was to investigate whether certain information quality dimensions are inherently more reliable than others, i.e. 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 and Strong, 1996), four high level categories that provided "comprehensive coverage of the multi-dimensional IQ construct" were empirically derived (Lee et al., 2002, p. 135). 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's et al. classification: 'intrinsic', 'contextual', and 'representational' information quality. We disregard '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 information quality 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 another; rather, we selected a subset of dimensions that others have argued for as to their importance, and investigated the more specific concern of possible differences in the consistency with which users could assess quality within this subset of four IQ dimensions chosen.

The Measurability of Information Quality Dimensions

To the best of our knowledge, no prior studies have explored the extent to which information quality 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. It must be emphasized that the concept of inter-rater reliability at the heart of this study is fundamentally different from the notion of construct validity that is regularly investigated in empirical studies of information quality, 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 et al., 1991). Two subtypes of construct validity are: convergent validity (the extent to which measures of the same construct are highly correlated) and 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 1-7 Likert scale, assume that Jack provides the ratings described on the left hand side of Table 1.

Insert Table 1 here

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 is also high, since the scores for ‘accuracy’ and ‘objectivity’ are clearly different. To illustrate how the notion of inter-rater 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 inter-rater reliability is quite different. The inter-rater 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 pre-requisite in the analysis of quality dimensions' inter-rater reliability, but it does not determine the inter-rater reliability scores. While construct validity is often analyzed in empirical studies of information quality (Lim, 2009; Flanagin & Metzger, 2003; Lee, et al., 2002), 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 inter-rater reliability (Moskal, 2000; LeBreton & Senter, 2008), to date very little is known of the extent to which existing scales of information quality lend themselves to consistent assessments by independent judges.

Although prior studies provide no explicit data indicating that a certain quality dimension is inherently more measurable or reliable than others, extant literature does imply this. Differences in inter-rater 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 inter-rater reliability would be lower for the 'intrinsic' IQ category where such cues are less apparent and where specialized knowledge may be required.

In addition to studying the measurability of the four IQ dimensions – accuracy, objectivity, completeness, and representation – we were also interested in studying the extent to which a composite (or gestalt') information quality (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 inter-rater reliability of information quality 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 inter-rater reliability between four information quality dimensions: accuracy, objectivity, completeness, representation. As an extension of this investigation, we also look at the reliability of an overall, or composite, information quality (CIQ) score.

Research Method

We employed a sample of 270 undergraduate student assessors that were recruited from a 3rd-year class at a North American university's Business School. The majority of students were in their early twenties, with a near even male-female distribution. In order to measure assessors' inter-rater 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 7-point Likert scale (from Strongly Disagree to Strongly Agree), using the set of items described in Table 2.

Insert Table 2 here

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, so as to represent the range of Wikipedia topics. We built on Wikipedia's top-level classification[4] (Kittur et al., 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, and thus we set an upper limit of 3500 words on article length, thus excluding lengthy outliers. Examples of Wikipedia articles included in this procedure are: Bricriu (troublemaker and poet in the Irish mythology), Dhol (a drum used in India), and Jacobi identity (in mathematics).

We employed a multi-step research design in order 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 Wikipedia articles from the set, resulting in 5-6 different assessments for each article (please refer to 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 *composite information quality (CIQ),* and represent their level of agreement with corresponding statements of quality on a 7-point Likert scale (from 'Strongly Disagree' to 'Strongly Agree'). This process ensured that inter-rater 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 1 and 6 ratings each. We dropped articles with very few (less than 3) ratings, arriving at a set of 98 articles, each having 3-6 assessments (see Step B in Figure 1). In order to ensure that inter-rater 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 ten sets, each containing 3 assessments on all Wikipedia articles. If an article was rated by 4, 5, or 6 students, then we would randomly select 3 student ratings for that article (see Step C in Figure 1). Inter-rater agreement was calculated independently for each of the ten sets, and we employed the average of these ten calculations in our analysis.

Insert Figure 1 here

In calculating inter-rater agreement, we first validated that the questionnaire items described in Table 2 indeed represented the information quality dimensions of interest and that construct validity was good (see details in Results section). 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 inter-rater reliability. Recall that inter-rater reliability (also referred to as 'inter-rater 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. Inter-rater reliability is often 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 et al., 2008). There are a number of statistics which can be used to determine inter-rater 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 kappa (Cohen, 1960), and its extension to more than two raters - Fleiss' kappa (Fleiss & Cohen, 1973) - take into account the amount of agreement that could be expected to occur through chance.

In this study we used the intra-class correlation (ICC) statistic (Haggard, 1958; Landis & Koch, 1977), which is directly analogous to Fleiss' Kappa (Fleiss & Cohen, 1973; Fleiss, 1981). Specifically, we used the intra-class 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 Alpha 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 4 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 intra-class agreement is mediocre (0.47) since absolute values differ.

Our method for calculating the statistical significance of differences in inter-rater reliability follows the approach employed by Klein et al. (2001) and Wong (2008), where the standard deviation is calculated for each of the items (in our case, Wikipedia article) that was rated by multiple assessors. We repeated this calculation for each of the IQ dimensions independently. We then used the assessments' standard deviation as an outcome variable, and tested the significance of differences in means using a paired-sample test (2-sided).

Results

Instrument Validation

To validate our measures, we conducted a principle component analysis (PCA) with varimax rotation using SPSS. It produced a four-factor solution, corresponding to the four quality dimensions we've 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 Composite Information Quality (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 were 0.26), suggesting that CIQ is indeed a higher-level construct that encompasses the dimensions of *Accuracy*, *Completeness*, and *Representation*. Interestingly, our data suggests 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, standard deviation, and factor loading of each measurement item.

Insert Table 3 here

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 alpha values 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 inter-correlations among the variables and their AVEs.

Insert Table 4 here

Inter-Rater Reliability Results

In analyzing inter-rater agreement of information quality measures, for each measure we used the average rating of the corresponding items, as illustrated in Table 5 below.

Insert Table 5 here

Generally speaking, we found that inter-rater agreement levels were low. Landis and Koch (1977) provide a scale for interpreting Kappa inter-rater value. A similar interpretation of ICC values was made by Fleiss (Fleiss & Cohen, 1973; Fleiss, 1981). Their 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, it should be noted 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 0.18-0.36 (see Table 5 above). However, these values still represent only moderate agreement, and were also likely influenced by the relatively large number of categories we employed.

When analyzing the differences in inter-rater reliability between the various quality dimensions, we notice 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 inter-rater agreement was assessed based on the standard deviations in raters’ assessments (using a paired-sample 2-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 inter-rater reliability for the *composite information quality* (*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 (the differences from *Completeness* and *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 paper we investigated the measurability of information quality 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 information quality dimensions: *accuracy*, *completeness*, *objectivity*, and *representation*. In order to explore the consistency in measurement of information quality dimensions, we performed a study of inter-rater 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 the variance for *Completeness*, we can say that *Objectivity* as an indicator is more difficult to measure than *Completeness*, and hence the measure is less reliable. Our findings indicate that there are substantial differences between inter-rater reliability scores for the different quality dimensions, such that there is less consistency in the ratings of some indicators compared to others.

The most striking finding from our study is that inter-rater reliability levels were very low, across all quality dimensions. Intra-class agreement levels did not exceed 0.17, indicating poor inter-rater reliability. For a comparison, a recent study of information literacy measurement tools, reports that in most cases the inter-rater reliability levels were in the 0.2-0.6 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 information quality 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 inter-rater reliability. The inter-rater 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 inter-rater 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, i.e. 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 (c.f. 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 information quality, '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: the assessment of some quality dimensions requires less domain expertise than 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 inter-rater reliability.

A secondary contribution of our study concerns the composite construct of information quality. While various studies refer to information quality 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 information quality incorporates 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 *composite information quality* (*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. It is interesting to note that the agreement levels for *CIQ* were higher than 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 information quality 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 inter-rater 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 inter-rater reliability with other constructs.

Conclusion

The notion of information quality is of primary concern to information science scholars, and it has attracted significant attention in recent years. Various conceptualization of IQ have been proposed, and most frameworks concur that information quality 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 information quality. Empirical studies of information quality often employ a survey to assess reader's 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 information quality 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 information quality 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 information quality 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 others - i.e. *Accuracy*, *Objectivity* - it is recommended that future studies of information quality 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 information quality 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 is also 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 information quality metrics for their published content. Often, these metrics are based on users' ratings. For example, many health-related web sites have tools for estimating the quality of web pages, and use symbols as 'award' or 'seal' to indicate high quality pages. These tools rarely report on the inter-rater 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.

Limitations and Future Research

Our study provides only preliminary findings regarding the measurability of information quality 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 cover 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. Also, 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 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 towards 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 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 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, but our study provides novel insights regarding the reliability of various information quality constructs that can be utilized in mitigating the less useful outcomes of this subjectivity. Still, additional research is required in order to validate our findings in alternative settings, expanding the scope of investigation, and to exploring the role of additional factors that affect variations in agreement levels. We hope that our study will open to 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 (SSHRC).

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 from: http://www.firstmonday.org/issues/issue11_11/chesney/

Cohen, J., 1960, A coefficient for agreement for nominal scales, *Education and Psychological Measurement*. Vol. 20, pp. 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 & 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 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: John Wiley, pp. 38–46

Fleiss J.L., (1986). *Reliability of measurement*. The design and analysis of clinical experiments. New York: John Wiley, 2-32.

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

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

Gagliardi, A. and 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 (BMJ),* 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, NY: Dryden Press Inc; 1958.

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. and Yang C.C., 2006, Conceptual Analysis of Parallel Corpus Collected from the Web, *Journal of the American Society for Information Science and Technology,* Vol. 57, Issue 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 *27th Annual CHI Conference on Humand Factors in Computing Systems (CHI 2009)* (Boston 2009).

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 Fourteenth International Joint Conference on Artificial Intelligence 2*(12): 1137–1143, (Morgan Kaufmann, San Mateo).

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: 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.

Lenhart, M., Simon M., & Graziano, A. (2001) The Internet and Education, *Pew Internet and American Life Project*, Retrieved 13 Jan 2010 from http://www.pewinternet.org/Reports/2001/The-Internet-and-Education.aspx

Lim, S. (2009). How and Why do College Students Use Wikipedia. *Journal of the American Society for Information Science & Technology*, *60*(11), 2189-2202.

Linstone, H. A., & Turoff, M. (Eds.). (1975). *The Delphi Method: Techniques and Applications.* Retrieved from http://is.njit.edu/pubs/delphibook/

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., and Hong C.K. (2008). Improving Wikipedia's Accuracy: Is Edit Age a Solution? *Journal of the American Society for Information Science & 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., Flanagin, A. J., & Zwarun, L. (2003). College student web use, perceptions of information credibility, and verification behavior. *Computers & Education, 41*(3), 271-290.

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 & Technology*, 60(5), 969-983.

Rainie, L. & Tancer, B. (2007). Wikipedia Users, *Pew Internet and American Life Project*, Retrieved 13 Jan 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. In Physical Therapy. Vol. 85, No. 3, pp. 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 13 Jan 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 AIS, 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.

Wallis, J. (2005). Cyberspace, information literacy, and the information society, *Library Review*, 54(4), 218-222.

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., and 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.

Footnotes

[1] It should be emphasized 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.

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

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

[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'.

Tables

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

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

Table 3.

Item means, standard deviations, and factor loadings.

Construct	Item	Mean	SD	Factor1	Factor2	Factor3	Factor4
Accuracy	Acc1	5.20	1.31	.**93**			
	Acc2	5.16	1.27	.**89**			
Completeness	Comp1	3.72	1.72		.**90**		
	Comp2	3.65	1.69		.**89**		
Objectivity	Obj1	5.20	1.51			.**85**	
	Obj2	5.21	1.51			.**89**	
Representation	Rep1	5.38	1.59				.**83**
	Rep2	5.40	1.31				.**80**
	Rep3	5.45	1.38				.**84**
CIQ	CIQ1	4.78	1.47	.56	.58		
	CIQ2	5.01	1.51	.34	.72		

Note: Factor loadings below 0.30 were suppressed.

Table 4.

Means, standard deviations, reliability, intercorrelations, and average variance extracted

	Construct	Mean	SD	α	1	2	3	4	5
1.	Accuracy	5.18	1.23	.91	.**96**				
2.	Completeness	3.68	1.62	.90	.36**	.**89**			
3.	Objectivity	5.20	1.35	.74	.26**	.24**	.**95**		
4.	Representation	5.41	1.22	.81	.31**	.38**	.27**	.**86**	
5.	CIQ	4.89	1.36	.80	.58**	.68**	.27**	.48**	.**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 2-tailed

Table 5

Inter-rater agreement results for the various constructs.

	Intra-Class	Variance			Common Inter-Item	Reliability
Users	Agreement				Correlation	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.

Figures

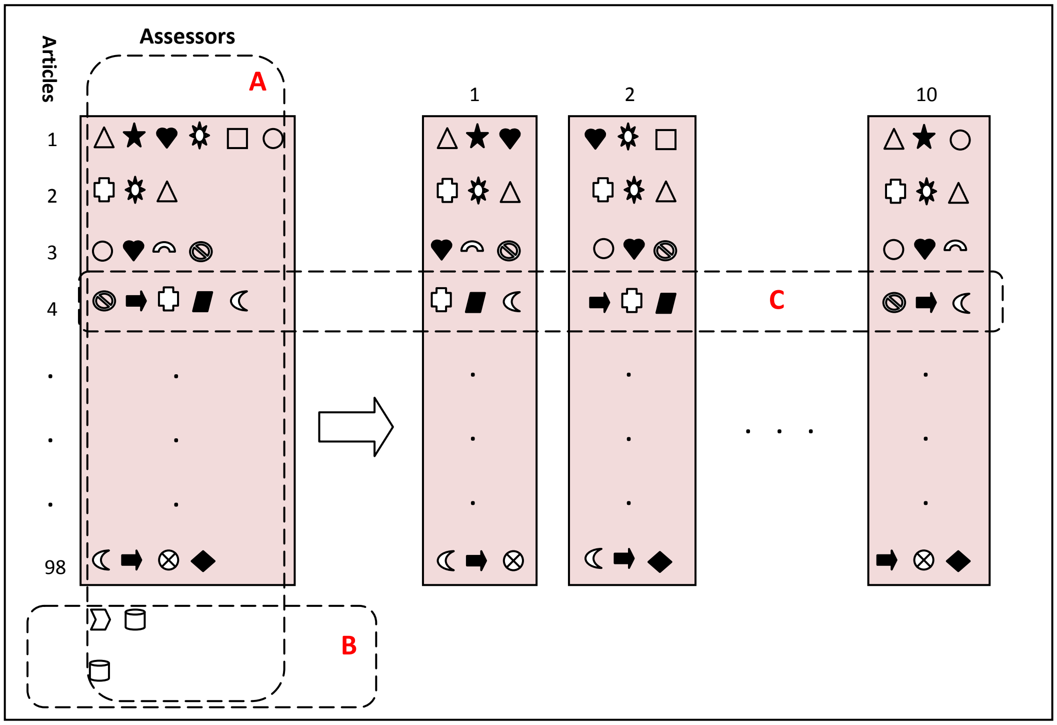

Figure 1. Sampling procedure. Assessors analyzed Wikipedia articles from a pre-defined set. Each of the distinct shapes represents a unique assessor. A large number of assessors were each assigned to two articles (Step A), and articles with less than three assessments were removed from the procedure,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