Qual Quant (2014) 48:2703–2718
DOI 10.1007/s11135-013-9919-0

On the choice of measures of reliability and validity in the content-analysis of texts

Anton Oleinik · Irina Popova · Svetlana Kirdina · Tatyana Shatalova

Received: 22 February 2013 / Accepted: 5 August 2013 / Published online: 7 September 2013

Abstract The paper discusses several reliability measures: Scott's pi, Krippendorff's alpha, free marginal adjustment (Bennett, Alpert and Goldstein's S), Cohen's kappa, and Perreault and Leigh's I and the assumptions on which they are based. It is suggested that correlation coefficients between, on one hand, the distribution of qualitative codes and, on the other hand, word co-occurrences and the distribution of the categories identified with the help of the dictionary based on substitution complement the other reliability measures. The paper shows that the choice of the reliability measure depends on the format of the text (stylistic versus rhetorical) and the type of reading (comprehension versus interpretation). Namely, Cohen's kappa and Bennett, Alpert and Goldstein's S emerge as reliability measures particularly suited for perspectival reading of rhetorical texts. Outcomes of the content analysis of 57 texts performed by four coders with the help of computer program *QDA Miner* inform the analysis.

Keywords Reliability measures · Content analysis · Correlation analysis · Interpretation · Comprehension · Stylistic texts · Rhetorical texts

A. Oleinik (✉)
Department of Sociology, Memorial University of Newfoundland, St. John's, NL A1B 5S7, Canada
e-mail: aoleynik@mun.ca

A. Oleinik
Central Economics and Mathematics Institute Russian Academy of Sciences, Moscow, Russia

I. Popova
Institute of Sociology Russian Academy of Sciences, Moscow, Russia

I. Popova
Higher School of Economics, Moscow, Russia

S. Kirdina
Institute of Economics Russian Academy of Sciences, Moscow, Russia

T. Shatalova
Moscow State Lomonossov University, Moscow, Russia

1 Introduction

The assessment of a study in terms of its validity and reliability represents an important step in research design, implementation and dissemination. The correspondence between, on one hand, a research instrument or research finding and, on the other hand, the phenomenon under study make them valid. Reliability depends on the stability of a research instrument or research finding over time and/or various scholars. There are several forms of reliability, depending on the epistemological stance that one takes (positivist or interpretativist) and on the characteristics of the research design: reliability–stability, reliability–reproducibility and reliability–accuracy.

Content analysis necessitates paying special attention to reliability–reproducibility. Content analysis is intended to describe, analyze and interpret the meanings that a text or image contains. The applications range from coding open-ended questions in mass surveys (Scott 1955; Muñoz-Leiva et al. 2006) to creating annotations in linguistics and library studies (Artstein and Poesio 2008). If a single researcher performs the tasks of reading a text and coding it with the help of a code book, the coder's judgments may be highly subjective. The involvement of multiple coders, however, calls for gauging the strength of agreement between them. Krippendorff (2004a, p. 215) defines reliability–reproducibility as "the degree to which a process can be replicated by different analysts working under varying conditions, at different locations, or using different but functionally equivalent instruments".

Content analysts have several measures of reliability–reproducibility at their disposal: Bennett, Alpert and Goldstein's S, Scott's π, Cohen's κ, Krippendorff's α, to cite just a few. This plurality creates some confusion: do these measures replace or complement one another? In the case of replacement, one simply needs to identify the best measure of reliability that "outperforms" the others. A claim that "of the existing measures, Krippendorff's α is best suited as a standard" (Hayes and Krippendorff 2007, p. 78) serves as an illustration of such approaches. In the case of complementarity, one attempts to identify the areas where the existing measures apply. Under which conditions (parameters of the texts subject to analysis, the number of coders, the number of categories included in the code book) is the use of Krippendorff's α more appropriate than the use of Bennett, Alpert and Goldstein's S?

This article is intended to provide additional arguments and empirical evidence to support the assumption of the complementary nature of the reliability measures (see also Artstein and Poesio 2008, pp. 586–591; Muñoz-Leiva et al. 2006, pp. 530–533). The purpose of the study reported below was to tentatively outline the areas of application for the most popular reliability measures in keeping with the particularities of the text subject to content analysis and of the specific objective of the content analysis (comprehension of the author's message as opposed to the reader's interpretation).

The article has three parts, along with an introduction and a conclusion. The first part provides a brief overview of the most popular reliability–reproducibility measures. The discussion of the assumptions on which each reliability measure is based serves to formulate a research question as to how the context of content analysis influences their applicability. The second part discusses the sources of the data, namely a study of academic reading (how scientists read and understand one another's works). In the third part, two hypotheses with respect to the choice of reliability measure are tested empirically.

2 Choosing reliability measures in function of the context of content analysis

2.1 Existing measures: an overview

A comprehensive overview of the existing reliability–reproducibility measures, including the discussion of the mathematics underpinning their calculation, can be found elsewhere (Krippendorff 2004a; Neuendorf 2002). Our task here is more limited. It involves the identification of differences in the assumptions on which the reliability measures are based.

Differences in the assumptions often remain overshadowed by the fact that most known reliability–reproducibility indexes derive from a common logic. Setting the simple percentage of agreement aside (percentage of units coded by all coders in the same manner), the calculation of the reliability measures involves comparing the observed percentage of agreement or disagreement with the expected one. In other words, the level of (dis)agreement achieved is compared to the level of (dis)agreement that could be obtained by chance. "The value $1 - A_{\mathrm{e}}$ [A stands for agreement] will then measure how much agreement over and above chance is attainable; the value $A_{\mathrm{o}} - A_{\mathrm{e}}$ will tell us how much agreement beyond chance was actually found. The ratio between $A_{\mathrm{o}} - A_{\mathrm{e}}$ and $1 - A_{\mathrm{e}}$ will then tell us which proportion of the possible agreement beyond chance was actually observed. This idea is expressed by the following formula: $S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e}$" (Artstein and Poesio 2008, p. 559). Krippendorff's α has a similar rationale, but it refers to the levels of disagreement, D_{o} observed and D_{e} expected (Krippendorff 2004a, p. 248).

The reliability measures start to diverge at the point of calculating the expected (dis)agreement, A_{e}. This apparent technicality, nevertheless, has its origin in the different assumptions on which these measures are based.

2.1.1 Scott's $\boldsymbol{\pi}$

Scott's π corrects the percentage of agreement "for the number of categories in the code, and the frequency with which each is used" (Scott 1955, p. 323). It is achieved by comparing the observed distribution of the categories with the expected one, as in the general case.

Scott makes a very specific assumption as to the correspondence between the observed and expected distribution, however. It is assumed that the observed patterns in the assignment of the categories by the coders are indicative of their "true" distribution, which serves as a basis for calculating the expected distribution. In other words, π uses "the actual behavior of the coders to estimate the prior distribution of the categories" (Artstein and Poesio 2008, p. 561).

The assumption made by Scott implies that the random (expected) assignment of the categories to the coded units (texts), by any coder, is governed by the distribution of the categories in the actual world. It follows that trained and competent coders normally identify all units that correspond to a particular category. If one coder overlooks a relevant unit, the second coder, who works independently from the first but uses the same code book, will probably spot the omitted unit and vice versa.[1] The units identified by the coders' joint efforts provide an approximate match of their prior, "true" distribution that exists independently of the coders' input.

The coders are interchangeable and equally qualified. The more coders are involved in the process of coding, the fewer chances there are of omissions. The original formula for

[1] A similar assumption underpins the use of Cronbach's α in the cultural consensus theory. The agreement between coders presumably depends on how well they know the content of a cultural domain that exists independently of their input (Weller 2007, p. 343).

calculating π in the case of two coders was subsequently generalized to the case of more than two coders by calculating π for each pair of coders and adding them up (Muñoz-Leiva et al. 2006, p. 526).

2.1.2 Krippendorff's $\boldsymbol{\alpha}$

π and Krippendorff's α derive from similar sets of assumptions: in the case of α parameters of the population (the "true" distribution of the categories) are guessed from the empirically observed proportions (from the margins of a contingency matrix). Krippendorff illustrates this point by considering an example of two coders, Jon and Han, who need to identify a particular category of articles (articles in Chinese with references to the USA). "With Jon finding US references in 2 out of 10 articles and Han finding them in 4 out of 10, the two observers have jointly identified 6 out of 20… this is our population estimate" (Krippendorff 2004a, p. 225). The parameters of the articles with references to the USA depend neither on whether they are subject to the content analysis nor on its outcomes.

This simple example clearly suggests: the "true" distribution of the categories pre-exists the coding process. The assumption that the population parameters are known in advance may be warranted in some areas of studies, namely in medicine, where qualified coders normally have an idea about the proportion of units that fit the description of a particular category (as in the case of identifying diseases on the basis of observed symptoms). The outcomes of the content-analysis of a sample of texts or images serve to unveil the parameters. The population estimate then enters into the calculation of the expected level of disagreement between the coders.

The empirically observed proportions represent a key component of the formula for calculating Krippendorff's $\alpha : \alpha = 1 - \frac{n-1}{n}(b + c)/2\bar{p}\bar{q}$, where n refers to the number of values used jointly by both coders ($n = 2N$, the number of units in the contingency matrix), $(b+c)$—to the proportion of disagreements—mismatches—between the coders (off-diagonal cells of the contingency matrix), $\bar{p}\bar{q}$—to the population estimates (margins of the contingency matrix) (Krippendorff 2004a, p. 248). The irrelevance of the number of coders and the number of categories for the calculation of the expected distribution means that these variables are not entered in an explicit manner.

Similarly to the case of π, the calculation of α implies the assumption that coders are interchangeable.[2] According to advocates of α, a good measure of reliability "should be (a) independent of the number of observers employed and (b) invariant to the permutation and selective participation of observers. Under these two conditions, agreement would not be biased by the individual identities and number of observers who happen to generate the data" (Hayes and Krippendorff 2007, p. 79). Some empirical evidence indeed supports the claim that the number of coders does not have an impact on α (Muñoz-Leiva et al. 2006, p. 530).

The same source, however, suggests that α is dependent on the number of categories included in the code book, which undermines its stability and precision as categories become more numerous. The other reported limitation of α refers to the situation of the prevalence of particular categories in data. When data are highly skewed, coders may agree on a high

[2] This assumption also serves to minimize the influence of the coders' values on the outcomes of content analysis. If the content analysis is not value-free, then the coders have fewer chances to agree on the distribution of the categories. The possibility theorem, which is applicable to choices guided by values, states that "for any method of deriving social choices by aggregating individual preference patterns which satisfies certain natural conditions, it is possible to find individual preference patterns which give rise to a social choice pattern which is not linear ordering" (Arrow 1950, p. 330).

proportion of items being indeed correct to a high degree, yet α remains low (Artstein and Poesio 2008, p. 573).

2.1.3 Cohen's κ

In contrast to π and α, the calculation of Cohen's κ involves using the other point of reference. The chance agreement is interpreted in terms of the consistency with which a coder categorizes units of analysis (Artstein and Poesio 2008, pp. 561, 570). From this point of view, the margins of the contingency table are indicative of the coders' individual preferences and biases (Perreault and Leigh 1989, p. 139), as opposed to the actual distribution of units among categories, as in the case of π and α. κ was originally calculated for the case of two coders. Its formula was subsequently generalized to the case of multiple coders by aggregating pairwise coefficients of agreement (Siegel and Castellan 1988, p. 285; Artstein and Poesio 2008, p. 562; Muñoz-Leiva et al. 2006; Dijkstra and van Eijnatten 2009, p. 763).

The underlying assumption is that "the probability that an object is assigned to a particular category does not vary across raters" (Siegel and Castellan 1988, p. 291). In other words, raters agree not because their choices are guided by an "invisible hand" of the "true" distribution of categories, but because they hold similar beliefs or have similar biases. Their choices tend to be subjective, which does not prevent them from reaching an agreement. This assumption appears warranted in paradigmatic sciences with psychology as a prime example among the social sciences. Thus, the observation that Cohen "was concerned mainly with psychological applications for which there often would be clearly established prior knowledge of the likely distribution of observations across cells" (Perreault and Leigh 1989, p. 139) comes as no surprise.

The change of the point of reference from the "true", naturally occurring distribution of categories to the coders' subjective judgments about it led to fierce criticism from advocates of π and α. According to them, "κ is concerned with the two individual observers, not with the population of data they are observing, which ultimately is the focus of reliability concerns" (Krippendorff 2004a, p. 248; see also Hayes and Krippendorff 2007, p. 81). Nevertheless, κ and α may refer to different aspects of reliability: the reliability of judgments as opposed to the reliability of data (Krippendorff consistently uses only the latter).

2.1.4 Bennett, Alpert and Goldstein's S

Bennett, Alpert and Goldstein's S represents another example of the reliability coefficient using the coder's subjective judgments as a point of reference. S "is sensitive to the number of categories available but says nothing about the population of data" (Krippendorff 2004a, p. 248). Compared with α and π, S sides with κ because their calculation involves making no assumptions about the "true" distribution of categories. Compared with κ, S gives less weight to the coders' individual preferences and biases. Instead, in the case of S the expected agreement between coders is calculated under the assumption of the equal probability of applying a particular category. If there are two categories and two coders, then their chances of agreeing are equal to $2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$ (chances of two independent events in two symmetrically positioned cells of the contingency matrix).

The rationale behind the assumption of the equal probability of applying a particular category could be better understood by looking at the context in which this reliability measure was initially proposed. Bennett and co-authors aimed to compare answers given by the respondents to questions on the same topic asked in the format of, on one hand, closed-ended questions in a mass survey and, on the other hand, semi-structured questions in an in-depth qualitative interview (Bennett et al. 1954, p. 305). No "true value" (true distribution of answers) exists in these circumstances because each method for data collection has its advantages and disadvantages. Arguably, a similar situation characterizes most non-paradigmatic sciences with no solid consensus among the scholars working in these fields. This line of reasoning may explain the high popularity of S: by Krippendorff's account, S has been reinvented in various forms at least five times since the mid-1950s (Krippendorff 2004a, p. 245; Hayes and Krippendorff 2007, p. 80), including in the form of Perreault and Leigh's I_r discussed in the next section.

In technical terms, the value of S depends on k, the number of categories (A_o stands for the observed agreement): $S = \frac{k}{k-1}(A_o - \frac{1}{k})$ (Bennett et al. 1954, p. 307). For this reason, it is subject to the following criticism: S tends to be inflated by the number of unused categories that the author of the instrument had imagined and by rarely used categories in the data (Hayes and Krippendorff 2007, p. 80). Nevertheless, this line of criticism assumes that an "ideal" code book (the one that perfectly matches the population) exists. It also omits the suggestion that α cannot be a perfect standard for assessing S's inflation because α tends to be deflated, especially when one category prevails.

The original formula for calculating Bennett, Alpert and Goldstein's S refers to the case of two coders and a nominal-level category with two attributes (e.g., Yes/No) only. S can be calculated in a general form as a sum of the distances from agreement between any number of coders that can be achieved by chance assuming equal probability of categories and their attributes: $S = 1 - \frac{1-\sum_{i=1}^{N}\frac{A_{o_i}-A_{e_i}}{1-A_{e_i}}}{N}$, where A_{o_i} refers to the observed agreement between a pair of coders when assessing unit i, A_{e_i}—to the expected agreement between them. $A_{e_i} = \frac{1}{2}$ in the case of a nominal category with two attributes, which serves to simplify the general formula: $S = 1 - \frac{2\sum_{i=1}^{N} A_{o_i}}{N} = 1 - \frac{2(b+c)}{N}$, where $(b + c)$ refers to the proportion of disagreements between the coders (off-diagonal cells of the contingency matrix).

A simulation using a dataset containing 12 units, one nominal category with two attributes and several coders (2, 3 and 4) shows that α does tend to be deflated (Fig. 1). Regardless of the number of units with a disagreement (this varies from 0 to 12 in the present case), α does not exceed 0, thereby suggesting a lower level of agreement than one that could achieved by

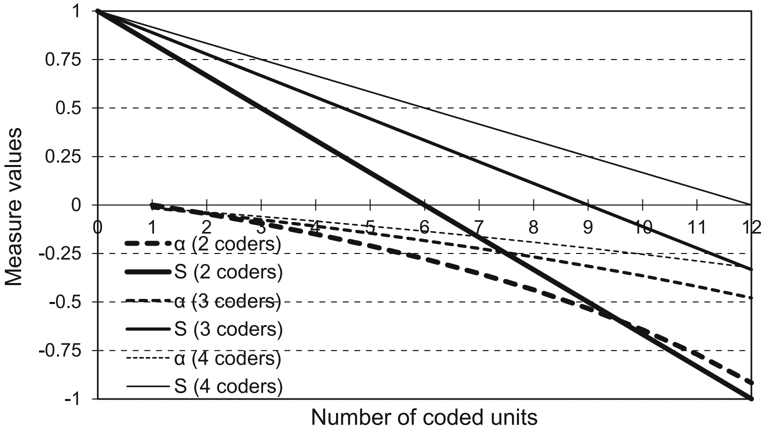

Fig. 1 Values of Krippendorff's α and Bennett, Alpert and Goldstein's S depending on the number of units with a disagreement; 2, 3 and 4 coders; 12 coded units

chance. S, on the other hand, varies from +1 to −1 (when two coders are involved) responding well to changes in the number of coders and the number of units with a disagreement. There was no other regularity in the distribution of categories, except with respect to the number of units with a disagreement.

2.1.5 Perreault and Leigh's $\boldsymbol{I}_r$

In numerical terms, I_{r} represents the square root of Bennett, Alpert and Goldstein's S (Muñoz-Leiva et al. 2006, p. 526). Perreault and Leigh use an algorithm for calculating I_{r} that significantly differs from Bennett, Alpert and Goldstein's formula, nevertheless. "Reliability here can be thought of as the percentage of the total responses (observations) that a typical judge could code consistently given the nature of the observations, the coding scheme, the category definitions, the directions, and the judge's ability" (Perreault and Leigh 1989, p. 140).

Described in this manner, I_{r} takes both objective and subjective factors into account. The authors place I_{r} half-way between two reference points, the population parameters and coders' individual characteristics (i.e. somewhere between α and κ). According to Perreault and Leigh, these assumptions fit the situation in marketing where, on one hand, the distribution of responses is not known in advance and, on the other hand, the coders aim to produce an objective picture by attempting to minimize the impact of individual preferences and biases.

2.1.6 Pearson's $\boldsymbol{r}$

The use of Pearson's correlation coefficient r in content analysis has led to controversy. Krippendorff disapproves of running correlations between the outcomes of the content analysis performed jointly by several coders (Krippendorff 2004a, p. 245). Namely, he argues that their use involves a circularity problem: the performance of one coder should not be assessed by referring to the performance of another. When coders do their job independently one from another, one's decisions cannot "cause" the other's choices in any meaningful manner.

Cronbach's α is the primary target of Krippendorff's criticisms. Cronbach's α is calculated from the number of people involved in the coding and the agreement among them: $\alpha = \frac{n\bar{r}}{1+(n-1)\bar{r}}$, where n is the number of coders being combined, and $\bar{r}$ is the average correlation coefficient r between all pairs of individuals (Weller 2007, p. 343). Cronbach's α can be calculated for quantitative categories measured at the ratio, interval and ordinal levels, which limits its applicability to content analysis. It is worth emphasizing that Cronbach's α has multiple applications beyond the calculation of the level of inter-coder agreement. For instance, Cronbach's α is also used as a measure of the internal consistency, or reliability, of the items constituting the scales for individual respondents (Camp et al. 1997). Krippendorff's arguments do not undermine the uses of Cronbach's α as a measure of the respondent's consistency.

There is no need to assume causality every time one performs a correlation analysis for the assessment of inter-coder agreement, however. The existence of an association between variables is a necessary but not sufficient condition for assuming a causal relationship between them (Bryman et al. 2012, p. 23). Furthermore, correlation could be performed not only between coders' outputs.

Other authors do not rule out the use of correlations in content analysis. In a useful comparison of R- and Q-mode research designs, Dijkstra and van Eijnatten (2009) observe that correlations are commonly used in the R-mode design with the purpose of analyzing the similarities between variables over research units. In content analysis, the text or image is the research unit whereas variables refer to various text/image parameters (source, genre, date, etc.), eventually including the names of the coders. The Q-mode, which is intended to study the resemblance between research units over variables does not exclude correlation analysis either. For instance, inter-class correlations could be performed when several coders use a large number of categories in the Q-mode (Dijkstra and van Eijnatten 2009, p. 764). Warner (2008, pp. 831–832) holds a similar opinion, adding that this only applies to quantitative categories measured at the ratio, interval and ordinal levels.

Oleinik (2010) proposes running correlations between content analysis outcomes and selected quantitative text parameters (for example, measures of word co-occurrence). This approach serves to avoid the circularity problem brought up by Krippendorff. It also paves the way for finding a mid-point between references to the "true value" (a text's inherent characteristics) and to the coders' interests, intentions and ability. In contrast to Cronbach's α, the correlational analysis proposed by Oleinik is not limited to quantitative categories only. It involves quantifying qualitative (nominal-level) categories. For instance, a qualitative coding distribution in a sample of texts/images is represented in the form of a matrix with cases (texts, images) in rows and qualitative codes in columns (their frequency or presence/absence). Then distances between vectors (cases) are expressed in quantitative terms (such as the Jaccard coefficient of similarity or the Cosine coefficient of similarity). The next section contains some arguments justifying the appropriateness and methodological soundness of this form of correlation analysis in the context of the content analysis of texts.

2.2 Particularities of the content analysis of texts

As opposed to the content analysis of open-ended questions and other naturally occurring units, unitizing is problematic in the content analysis of texts and images (Artstein and Poesio 2008, p. 580; Oleinik 2010, p. 872). Unitizing involves identifying contiguous sections containing information relevant to a research question within a medium (text, image, video recording) (Krippendorff 2004a, p. 219). "For reliability to be perfect, the units that different observers identify must occupy the same locations in the continuum and be assigned to identical categories" (Krippendorff 2004b, p. 792). To the best of our knowledge, the existing computer programs do not offer the option of calculating reliability measures for unitizing, which deflates all reliability measures for coding texts and images. The reliability of unitizing tends to be a confounding variable when assessing the reliability of coding and identifying the latter's factors such as the clarity of coding instructions for coders.

In addition to technical particularities, the content analysis of texts also has more substantial characteristics. A text as a communication medium may have multiple meanings. When writing a text, its author conveys a message. When reading the text, its readers—in the plural, since texts normally have multiple readers—may attempt to "decipher" the author's message. The readers may as well bring new perspectives to the text, discovering new ideas in it depending of their specific background, interests, values and ability. In this regard, Norris and Philips differentiate between comprehension and interpretation (Fig. 2). "In comprehension, the person is aiming to grasp the meaning of something that has that meaning in it, as it were, perhaps as the author's intention or as the design of a higher being if the task is comprehending the world. In interpretation, the person is placing meaning on something that may or may not be the meaning inherent in the object" (Norris and Philips 1994, p. 402).

Text formats necessitates placing greater emphasis on either interpretation or comprehension. Two text formats are differentiated in semiotics: stylistic and rhetorical (Lotman 1990, pp. 45–51). Rhetorical texts (novel, poem, diary, or essay) have a loose structure.

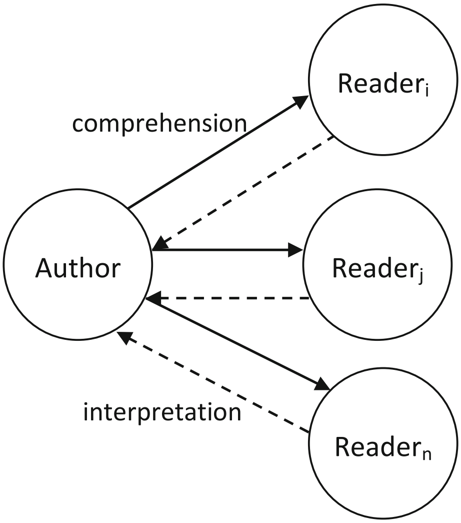

Fig. 2 The content analysis of texts as a set of interactions between the author and the reader

Metaphors and analogies abound in such texts. The content analysis of rhetorical texts calls for prioritizing interpretation: after all, they aim at provoking free associations and creative thinking.

Stylistic texts (scholarly article, textbook, or scientific letter) have a clear, often rigid structure. Arguments in stylistic texts must meet high logical standards: exhaustiveness and mutual exclusiveness of categories, transitivity and consistency in their rank ordering and so forth. Comprehension seems to be more appropriate for the content analysis of stylistic texts as a result of their orientation toward conveying a message in the least ambiguous manner. Interpretation of stylistic texts and comprehension of rhetorical texts are not excluded either, being special cases.

Thus, the content analysis of texts shall, on the one hand, be placed in the context of interactions between the author and the readers and, on the other hand, take into account the characteristics of the texts under study. Krippendorff acknowledges that a text can be read and, consequently, content analyzed in several manners. Namely, he discusses three alternative assumptions underlying content analysis: content is (i) inherent in a text, (ii) a property of the source of a text, and (iii) emerging in the process of a researcher analyzing a text relative to a particular context (Krippendorff 2004a, p. 19). He then argues that only the third assumption holds, however: “it is a major epistemological mistake to assume that texts have inherent meanings or speak for themselves” (Krippendorff 2004b, p. 789). The deliberate restriction of content analysis to perspectival reading makes coding less indeterminate at the price of excluding the case of comprehension from its scope. It also creates an epistemological inconsistency in the perception of various reliability measures: often criticized as being “too subjective” κ and S correspond better to the spirit of perspective-relative coding than the more “positivistic” π and Krippendorff’s α.

The objective of the present study, as stated in the introduction, can now be discussed in concrete terms. This article aims to explore the question as to which is the most appropriate reliability measure depending on text format (stylistic versus rhetorical) and the type of reading (comprehension versus interpretation). For this purpose, it would be sufficient to focus on the reliability measures applicable to (i) nominal dichotomous categories (they are most commonly used in the content analysis of texts), (ii) multiple coders (because a text potentially has an unlimited number of readers),[3] and (iii) different reference points (the author's message as opposed to the reader's perspective).

The content analysis of stylistic texts involves choosing the author's message as a point of reference, which suggests a preference for π and α to κ and S. The author's message determines the parameters of the "true value" (the distribution of categories corresponding to the author's intentions). Coders agree when they correctly comprehend what the author wanted to say with the help of their code book and its consistent application in the process of qualitative coding.

When sending a "message" in the format of a stylistic text, the choice of words determines the scope of the possible for the author. Under this assumption, "the study of the range of things that speakers are capable of doing in (and by) the use of words and sentences" (Skinner 2002, p. 3) becomes possible and justifiable. Measures of word co-occurrence (e.g., the Cosine coefficient of similarity between texts, see Salton and McGill 1983, p. 203) provide the content analyst with a rough idea of the extent of overlapping areas between the scopes of the possible in various texts. In other words, measures of word co-occurrence may serve as a proxy for relative parameters of the "true value". The existence of a correlation between the distribution of qualitative codes in texts (as measured, for instance, with the help of the Cosine coefficient of similarity or the Jaccard coefficient of similarity) and word co-occurrence in these same texts suggests a reliable and valid character of the qualitative coding of stylistic texts (Oleinik 2010, pp. 868–871). The concept of validity is used here in the sense of a correspondence between what the author intends to convey to the reader and what the reader learns.

The reader's perspective prevails in the content analysis of rhetorical texts, which explains the opposite order of preferences: κ and S rather than π and α. Arguably, the assumption of the interchangeability of coders does not hold in the content analysis of rhetorical texts. Each new reader brings a new perspective with respect to the text. Consequently, the identification of the "true value" becomes less relevant here than the agreement of coders on what seems important to them. Coders may produce a reliable coding scheme even if their code book does not match the categories that the author had in mind when writing the text. For this reason, measures of word co-occurrence are not of great assistance in assessing the reliability of the coding of rhetorical texts. Instead of looking at co-occurrences of all words, the content analyst may decide to classify them and to create a dictionary based on substitution with categories that match the code book for qualitative coding. The existence of a correlation between the distribution of qualitative codes in texts and the distribution of categories identified with the help of the dictionary based on substitution in these same texts suggests a reliable and valid character of the qualitative coding of rhetorical texts (Oleinik 2010, pp. 868–871). The concept of validity is used here in the sense of a correspondence between what the reader wishes to learn from the text and what the reader really gets from it.

This line of reasoning suggests a tentative classification of the reliability measures as a function of the format of the text and the type of reading (Table 1). The reliability of the content analysis based on comprehension proper to stylistic texts can be better measured with the help of π, α, and Pearson's r (between word co-occurrence and qualitative coding). The reliability of the content analysis based on interpretation proper to rhetorical texts can be better measured with the help of S, κ, and r (between dictionary based on substitution

[3] This rationale should not be confused with another, more "positivist" argument advanced by Krippendorff (2004a, p. 249): "we must estimate the distribution of categories in the population of phenomena from the judgments of as many observers as possible (at least two), making the common assumption that observer differences wash out in their average".

Table 1 Reliability measures appropriate for particular types of text and research tasks

Type of reading	Text format	
	Stylistic	Rhetorical
Comprehension	Scott's π, Krippendorff's α, Pearson's r (between word co-interpretation occurrence and qualitative coding)	Perreault and Leigh's I_r
Interpretation	Perreault and Leigh's I_r	Bennett, Alpert and Goldstein'sS, Cohen's κ, Pearson's r (between dictionary based on substitution and qualitative coding)

and qualitative coding). Finally, in two special cases, interpretation of stylistic texts and comprehension of rhetorical texts, I_r seems to be the most suitable because of an intermediate position of this measure on the scale ranging from the population parameters to the coders' individual characteristics.

2.3 Sources of the data

A comprehensive answer to the research question formulated above requires the content analysis of two sets of texts, stylistic and rhetorical. At this stage, we are able to report the outcomes of the content analysis of a sample of stylistic texts (scholarly articles, book reviews and essays) using a computer program for qualitative and quantitative content analysis, namely *QDA Miner* version 4.0.4 with module *WordStat* version 6.1.4. Its design serves, nevertheless, to test two specific hypotheses deriving from the research question. First, correlation coefficient r between, on one hand, the distribution of qualitative codes in texts and, on the other hand, word co-occurrence in these same texts can be used to assess the reliability and validity of the content analysis of stylistic texts. Namely, this correlation coefficient helps ensure that the qualitative coding matches the text's "true value" (the distribution of categories corresponding to the author's intentions). Second, compared with α, S tends to be more strongly associated with the distribution of qualitative codes, which derives from S's assumed relevance to perspectival reading. Unfortunately, *QDA Miner* cannot be used to calculate either κ or I_r, which would be even more appropriate under the circumstances.

Within the framework of a research project on academic reading, the authors of this present study content analysed a sample of their own scholarly publications. It must be noted that the co-authors did not have previous any experience of scientific collaboration and worked independently in the past. In total, 57 texts published between 1999 and 2011 were included in the sample (20 written by the first co-author, a sociologist and economist, 20—by the second, also a sociologist and economist, and 17—by the third, a sociologist).[4] The coders' task was to comprehend the texts rather than to interpret them. The presence of several scholarly essays (as opposed to scholarly articles) in the sample meant that some elements of interpretation could not be avoided, however.

[4] Being a recent university graduate, the fourth co-author has not produced enough publications yet. She played the role of a "perfect reader" whose take on a text is not affected by the authorship of the other texts included in the sample.

The content analysis was performed in several stages. At the first stage, the coders read the texts and developed their own code books for qualitative coding independently from one another. They were provided with uniform written instructions and used the same computer program. The reliability and validity of the qualitative coding was assessed with the help of Pearson's r calculated between, on one hand, the distribution of qualitative codes in texts and, on the other hand, word co-occurrence and the distribution of categories identified with the help of dictionaries based on substitution in these same texts. We pursued our research only after achieving a moderately strong level of correlation in each case.

At the second stage, the coders developed three common code books (one adapted to the first co-author's texts, the other—to the second co-author's texts and so forth) as well as three common dictionaries based on substitution containing the same categories and subsequently applied them when re-reading the texts independently from one another. This time, in addition to running the correlations (as during the first stage) we also calculated α. We proceeded only after achieving a moderately strong level of correlation in each case and an inter-coder agreement exceeding 0.5 for each pair of coders.[5]

Finally, at the third stage the three code books were merged. The same applies to the common dictionary based on substitution. The coders used the merged code book when coding all the texts. In other words, they attempted to identify fragments relevant to categories specific to a particular author not only in this author's texts, but also in the texts written by the other authors. We ran the correlations and calculated the inter-coder agreement coefficients, α and S. In the following, we report the results of the content analysis at the second and third stages. The findings of the first stage are discussed elsewhere (Oleinik et al. 2013). The application of the common code book in its two versions enabled us to assess the reliability of both comprehension (at the second stage) and interpretation (at the third stage).

2.3.1 Empirical test: assessing the reliability of the content analysis of scholarly texts

The outcomes of the correlational analysis suggest that the four coders were consistent when applying both versions of the code book. Namely, the distribution of qualitative codes is substantially correlated with word co-occurrence (Table 2). Arguably, this means that the distribution of categories corresponds to the author's intentions: the latter are reflected in the selection of words and their combinations. The values of r turned out to be higher than in a previous study carried out by the first co-author using a similar approach. He content analysed a sample of transcripts of semi-structured interviews—they contain more rhetorical elements than scholarly texts—and found r between the distribution of qualitative codes and word co-occurrence to vary from 0.295 to 0.483 (Oleinik 2010, p. 869). An increase in the level of correlation between the distribution of qualitative codes and the distribution of categories identified with the help of the dictionary based on substitution at the third stage may be indicative of a stronger emphasis on interpretation (i.e. on perspectival reading).[6]

In order to produce additional evidence in support of our claim that $r$should be added to the list of the reliability measures proper to the content analysis of texts, we compared its values with the values of α and S (Table 3). This time, we studied correlations between the distributions of the qualitative codes produced by each coder: how similar are they (using the Jaccard coefficient of similarity to measure the distance between a particular case and a

[5] When assessing this level of inter-coder agreement, one has to bear in mind that it reflects both the reliability of unitizing and the reliability of coding.

[6] The dictionary based on substitution was subject to small edits only at the third stage.

Table 2 Values of Pearson's correlation coefficients *r* between the distribution of qualitative codes in texts, word co-occurrence and the distribution of categories identified with the help of the dictionary based on substitution in these same texts, 4 coders, $N = 57$

	Stage II			Stage III		
	Qualitative coding	Dictionary based on substitution	Word co-occurrence	Qualitative coding	Dictionary based on substitution	Word co-occurrence
Stage II						
Qualitative coding	1	0.747**	0.674**	–	–	–
Dictionary based on substitution	0.747**	1	0.707**	–	–	–
Word co-occurrence	0.674**	0.707**	1	–	–	–
Stage III						
Qualitative coding	–	–	–	1	0.902**	0.684**
Dictionary based on substitution	–	–	–	0.902**	1	0.707**
Word co-occurrence	–	–	–	0.684**	0.707**	1

**Correlation is significant at the 0.001 level (2-tailed). The level of statistical significance is reported for the sake of comparison only: the sample was not random

Table 3 Values of Krippendorff's α, Bennett, Alpert and Goldstein's S, Pearson's correlation coefficients r between the distribution of qualitative codes

Coders	Measures						
	Stage II			Stage III			
	$\alpha(N=37)$	$S(N=37)$	$r(N=57)$	$\alpha(N=37)$	$S(N=37)$	$r(N=37)$	$r(N=57)$
A+S	0.575	0.820	0.908	0.412	0.640	0.818	0.774
A+I	0.535	0.802	0.949	0.436	0.679	0.934	0.909
A+T	0.555	0.812	0.869	0.434	0.698	0.957	0.674
S+I	0.544	0.813	0.833	0.423	0.653	0.848	0.656
S+T	0.519	0.811	0.755	0.404	0.666	0.873	0.534
I+T	0.496	0.797	0.831	0.399	0.687	0.937	0.721
Means for pairs	0.537	0.809	0.858	0.418	0.671	0.895	0.711
A+S+I	0.465	0.590	–	0.289	0.365	–	–
A+S+T	0.465	0.590	–	0.286	0.375	–	–
S+T+I	0.432	0.574	–	0.278	0.371	–	–
A+I+T	0.440	0.570	–	0.299	0.398	–	–
Means for triads	0.4505	0.581	–	0.288	0.377	–	–
A+S+T+I	0.367	0.420	–	0.208	0.245	–	–

N=57 refers to the number of texts; *N*=37 refers to the number of codes in the common code books; A, S, I and T refers to the coders' first names

centroid, the case occupying the central position on the 2D map of coding co-occurrences)? We used the code as a unit of analysis when running the correlations at the third stage. It corresponds to the logic of the Q-mode, as opposed to the R-mode, for which the text is a more natural unit of analysis. The common codebooks contained 37 codes.[7]

The use of the code as a unit of analysis when running the correlations at the second stage did not make much sense, however: it would have produced three completely disconnected clusters of codes. Thus, we used the text as a unit of analysis when running the correlations at the second stage: how similar (in terms of the Cosine coefficient of similarity) are distributions of cases in terms of coding co-occurrences obtained by each coder? This created some methodological impurity. But the two versions of the correlations performed at the third stage (one using the code as a unit of analysis and the other using the text as a unit of analysis) produced convergent rather than divergent outcomes.

The other limitation refers to the applicability of correlation analysis to pairs of coders only. Bivariate correlations cannot be used directly to assess the agreement between three and more coders. In the case of triads, quads and so forth, only pair comparisons can be made, which explains why some cells of Table 3 were left empty.

α turned to be moderately associated with r at the second stage: the coefficient of correlations between them was equal to 0.503 ($N = 6$ pairs).[8] The relationships between the reliability measures at the third stage need to be discussed in greater detail. α was still associated with r (coefficient of correlation was equal to 0.474 when using the text as a unit of analysis and 0.329 when using the code as a unit of analysis, $N = 6$ pairs in both cases).

[7] The code book for analyzing the first co-author's texts contained 13 codes, in the case of the second co-author it contained 15 codes, and in the case of the third—nine codes.

[8] The distribution of the reliability measures was visually inspected prior to correlation analysis. This "eye-balling" suggested that the normality of distribution condition was not significantly violated.

Thus, our first hypothesis can be tentatively accepted; at least, the correlation coefficients do not contradict the indications of the other reliability measures, namely α.

More substantially, S turned to be almost perfectly associated with r: coefficient of correlation was equal to 0.985 when using the code as a unit of analysis. This finding suggests that S and α may indeed refer to different aspects of the agreement between coders. S tends to be more strongly associated with the distribution of qualitative codes, which derives from S's assumed relevance to perspectival reading. It should be noted that the design of the third stage placed greater emphasis on interpretation as opposed to comprehension. S was not associated with r at the second stage (the correlation coefficient was -0.013). Thus, our second hypothesis can be tentatively accepted.

Table 3 also shows that the values of α depended on the number of coders. The mean for pairs of coders was higher than the mean for triads. The mean of triad was higher than α calculated for four coders. This finding undermines the assumption of the interchangeability of coders in the content analysis of texts, especially rhetorical texts (the differences in the means were particularly significant at the third stage).

3 Conclusion

The limited character of our sample prevents us from making far-reaching claims. However, our findings support rather than undermine our assumption that correlation coefficients can be used legitimately in the content analysis of texts. They complement the other reliability measures, such as Krippendorff's α and Bennett, Alpert and Goldstein's S. The results of our study also suggest that the choice of the reliability measure depends on the format of the text (stylistic versus rhetorical) and the type of reading (comprehension versus interpretation). Namely, S and Cohen's κ emerge as reliability measures particularly suited for the perspectival reading of rhetorical texts.

There are several directions for further developments. First, a more careful test of the hypothesis that the choice of the reliability measure depends on the format of the text and the type of reading involves comparing the outcomes of the interpretation of ideal typical rhetorical texts (poems, novels) and the results of the comprehension of ideal typical stylistic texts (scholarly articles in the paradigmatic sciences).

Second, on a methodological level, more efforts could be devoted to perfecting methods for calculating S and its variants. Krippendorff (2004a, pp. 223–249) discusses at least four methods for calculating α and, accordingly, its four interpretations. Artstein and Poesio (2008, pp. 565–566) use the sum of the square differences from the mean to calculate α. Arguably, a similar approach could be applied to the case of S. This would help address the most common criticism of S, namely its dependence on the number of unused codes.

Third, the reliability measures adapted to perspectival reading may have new, sometimes unexpected areas of applications. They are useful in any situation in which an assessor's subjective perception is important. Decisions made by panels of judges (i.e., by appellate courts) refer to one. If there is a disagreement between three or five judges constituting a panel, they often provide separate reasons for the same decision. The content analysis of such documents using π or α makes little sense: like readers of a rhetorical text, judges are not interchangeable. Thus, S and correlation coefficients may appear to be a better measure for the agreement of judges.

Fourth, in contrast to the content analysis of rhetorical texts, the reading of stylistic texts can be computerized more easily along the lines suggested in our paper (namely, using measures of word co-occurrence). This paves the way for adding new options to the databases of scholarly publications, such as the *Web of Knowledge*. One option refers to the eventual categorization of scholarly texts using not only a rather limited number of key words, but also word co-occurrences in general.

Acknowledgments The authors would like to thank the anonymous reviewers of *Quality & Quantity* for their helpful and constructive suggestions and comments. However, all remaining errors and inaccuracies are solely attributable to the authors.

References

Arrow, K.J.: A difficulty in the concept of social welfare. J. Polit. Econ. **58**(4), 328–346 (1950)

Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistic. Comput. Linguist. **34**(4), 555–596 (2008)

Bennett, E., Alpert, R., Goldstein, A.C.: Communications through limited-response questioning. Public Opin. Quart. **18**(3), 303–308 (1954)

Bryman, A., Bell, E., Teevan, J.J.: Social Research Methods, 3rd edn. Oxford University Press, Don Mills (2012)

Camp, S.D., Saylor, W.G., Harer, M.D.: Aggregating individual-level evaluations of the organizational social climate: a multilevel investigation of the work environment at the Federal bureau of prisons. Justice Q. **14**(4), 739–762 (1997)

Dijkstra, L., van Eijnatten, F.M.: Agreement and consensus in a Q-mode research design: an empirical comparison of measures, and an application. Qual. Quant. **43**(5), 757–771 (2009)

Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. **1**(1), 77–89 (2007)

Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. SAGE, Thousand Oaks (2004a)

Krippendorff, K.: Measuring the reliability of qualitative text analysis data. Qual. Quant. **38**(6), 787–800 (2004b)

Lotman, Y.: Universe of the Mind: A Semiotic Theory of Culture. Indiana University Press, Bloomington (1990)

Muñoz-Leiva, F., Montoro-Ríos, F.J., Luque-Martínez, T.: Assessment of interjudge reliability in the open-ended questions coding process. Qual. Quant. **40**(4), 519–537 (2006)

Neuendorf, K.A.: The Content Analysis Guidebook. SAGE, Thousand Oaks (2002)

Norris, S.P., Philips, L.M.: The relevance of a reader's knowledge within a perspectival view of reading. J. Read. Behav. **26**(4), 391–412 (1994)

Oleinik, A.: Mixing quantitative and qualitative content analysis: triangulation at work. Qual. Quant. **45**(4), 859–873 (2010)

Oleinik, A., Kirdina S., Popova I., Shatalova T.: Kak uchenye chitayut drug druga: osnova teorii akademicheskogo chteniya [How scientists read: on a theory of academic reading]. SOCIS 8 (2013)

Perreault, W.D., Leigh, L.E.: Reliability of nominal data based on qualitative judgments. J. Mark. Res. **26**(2), 135–148 (1989)

Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill Book Co., New York (1983)

Scott, W.A.: Reliability of content analysis: the case of nominal scale coding. Public Opin. Q. **19**(3), 321–325 (1955)

Siegel, S., Castellan, N.J.: Nonparametric Statistics for the Behavioural Sciences. McGraw Hill, New York (1988)

Skinner, Q.: Visions of Politics. Cambridge University Press, Cambridge (2002)

Warner, R.M.: Applied Statistics. SAGE, Thousand Oaks (2008)

Weller, S.C.: Cultural consensus theory: applications and frequently asked questions. Field Methods **19**(4), 339–368 (2007)