SC|M

Studies in Communication and Media

FULL PAPER

Coding quality in manual content analysis: An exploration of coder characteristics and category types for crowdworkers and student coders

Codierqualität in der manuellen Inhaltsanalyse: Eine Untersuchung der Codierendeneigenschaften und Kategorietypen bei Crowdworkern und studentischen Codierenden

Julia Niemann-Lenz, Anja Dittrich & Jule Scheper

Julia Niemann-Lenz (Dr.), University of Hamburg, Department of Social Sciences, Journalism and Mass Communication, Sedanstraße 19, 20146 Hamburg, Germany. Contact: julia.niemann-lenz(at)uni-hamburg.de, ORCID: https://orcid.org/0000-0003-1991-8535
Anja Dittrich (M.A.), Hanover University of Music, Drama & Media, Department of Journalism and Communication Research, Expo Plaza 12, 30539 Hanover, Germany. Contact: anja.dittrich(at)ijk.hmtm-hannover.de, ORCID: https://orcid.org/0000-0003-3766-5877
Jule Scheper (Dr.), Hanover University of Music, Drama & Media, Department of Journalism and Communication Research, Expo Plaza 12, 30539 Hanover, Germany. Contact: jule.scheper(at)ijk.hmtm-hannover.de, ORCID: https://orcid.org/0000-0001-6316-4238

FULL PAPER

Coding quality in manual content analysis: An exploration of coder characteristics and category types for crowdworkers and student coders

Codierqualität in der manuellen Inhaltsanalyse: Eine Untersuchung der Codierendeneigenschaften und Kategorietypen bei Crowdworkern und studentischen Codierenden

Julia Niemann-Lenz, Anja Dittrich & Jule Scheper

Abstract: Although the quantitative content analysis is one of the most important methods in empirical social research, the coding process receives little attention. This is particularly concerning in light of current developments, such as the rise of the use of crowdworkers as coders. Therefore, this study aims to shed light on the coding process by examining a) how the coding quality in terms of reliability and validity differs between student coders, which are traditionally often used for coding, and increasingly used crowdworkers, b) how coder characteristics such as personality traits and sociodemographics influence the coding quality, and c) how manifest and latent category types, that are coded with varying levels of difficulty for coders, impact coding quality. To test these research questions, 300 tweets on the topic of abortion were coded by both students and crowdworkers. A descriptive comparison reveals that the validity in both coder groups is sufficient for manifest, i.e. simple, category types but inadequate for latent, i.e. difficult, category types. Regarding reliability, student coders outperform crowdworkers slightly, particularly when more stringent criteria such as Krippendorff's alpha are considered. The results show that the coder characteristics have only a minor impact on quality, while category types have a significant impact.

Keywords: Content Analysis, coding, annotation, coder characteristics, crowdworker, reliability, validity, method research.

Zusammenfassung: Obwohl die Inhaltsanalyse eine der wichtigsten Methoden der empirischen Sozialforschung ist, findet der Codier-Prozess bislang nur wenig Beachtung. Dies ist besonders gravierend angesichts aktueller Entwicklungen, wie dem vermehrten Einsatz von Crowdworker*innen als Codierer*innen. Die vorliegende Studie zielt daher darauf ab, den Codierprozess genauer zu beleuchten. Konkret wird untersucht, wie a) die Codierqualität hinsichtlich Reliabilität und Validität zwischen studentischen Codierer*innen, die häufig zum Codieren genutzt werden, und neuerdings sehr beliebten Crowdworker*innen variiert, b) Merkmale der Codierer*innen, konkret Persönlichkeitsmerkmale und Soziodemografika, die Codierqualität beeinflussen und c) manifeste und latente Kategorietypen, die für Codierer*innen unterschiedlich schwer zu codieren sind, die Qualität der Codierung beeinflussen. Zur Überprüfung der Forschungsfragen wurden von Studierenden und

Crowdworker*innen Tweets zum Thema Abtreibung codiert. Ein deskriptiver Vergleich offenbart, dass die Validität in beiden Codierer*innengruppen für manifeste, also einfache, Kategorietypen ausreichend, für latente, also schwierige, Kategorietypen jedoch unzureichend ist. In Bezug auf die Reliabilität schneiden die studentischen Codierer*innen etwas besser ab als die Crowdworker*innen, insbesondere wenn strengere Kriterien wie Krippendorffs Alpha angelegt werden. Die Ergebnisse zeigen darüber hinaus, dass die Eigenschaften der Codierer*innen nur einen geringen Einfluss auf die Codierqualität haben, während die Kategorietypen einen signifikanten Einfluss haben.

Schlagwörter: Inhaltsanalyse, Codierung, Annotation, Codierer*inneneigenschaften, Crowdworker*innen, Reliabilität, Validität, Methodenforschung.

1. Introduction

Social and especially communication studies analyses often focus on media content in their investigation. The relevance of quantitative content analyses goes beyond a mere description of content because it is assumed that media produces certain effects and reflects the context and process of its production (Riffe et al., 2005, p. 22). While numerous textbooks are concerned with the methodology of content analysis (e.g., Früh, 2017; Krippendorff, 2013; Merten, 1995; Neuendorf, 2017; Rössler, 2017) and some studies address features such as reliability test documentation practices (e.g., Kolbe & Burnett, 1991; Lauf, 2001; Lombard et al., 2002; Lovejoy et al., 2014, 2016; Vogelgesang & Scharkow, 2012), the coding process itself represents a "black box" (Wirth, 2001, p. 158). A reason could be attributed to the limited influence that researchers have over this specific aspect of the analysis. However, it is the coding process that plays an important role in the application of content analysis; it is during this stage that data is generated, on the basis of which hypotheses or theories are subsequently confirmed or refuted (Wirth, 2001, p. 157). Data collection is the responsibility of coders who act in a rule-based procedure and are trained according to their coding tasks, but "coders are humans even when they are asked to act like computers" (Krippendorff, 2013, p. 127). It was already noted early on that differences in coding results should be investigated more closely, and that results could be related to coders' personalities (Spiegelman et al., 1953, p. 186). This, in turn, may have implications for reliability and validity of content analysis results (Wirth et al., 2015).

In recent years, it has become more and more common to use crowdworkers as coders (e.g., Benoit et al. 2016; Boxman-Shabtai, 2021; Budak et al., 2016; Hornik et al., 2022). Crowdworkers are individuals recruited to complete micro tasks for a small fee. A typical task is the labeling of texts and visuals, also referred to as annotating. Annotating and coding share certain similarities, although coding in media content analysis typically involves classifying media content into various categories and may also refer to different aspects and levels of the content. It may, therefore, be a more complex task. Comparisons of crowdworkers vs. student coders in quantitative content analysis report that both produce acceptable results, especially for coding simpler concepts (Atteveldt et al., 2021; Lind et al., 2017). Albeit the performance of trained student coders is considered superior, under appropriate conditions (illustrative codebook, training exercises) crowdworkers can achieve satisfactory performance even for more nuanced concepts (Budak et al., 2021; Lind et al., 2017).

In light of these promising results, the use of crowdworkers in manual content analysis can revolutionize the field by making data analysis more accessible, cost-effective and even more representative, as crowdcoding allows for recruiting a more diverse sample of coders. On the downside, coding with the crowd also presents challenges, such as ensuring the coding quality, addressing coder training needs, maintaining consistency across the coding team and ethical considerations. On average, crowdworkers earn less than $6 per hour according to a recent meta-study (Hornuf & Vrankar, 2022, p. 653). Low wages not only raise questions on research ethics and about appropriate compensation for research work; paying lower wages to crowdworkers leads to less time spent on the tasks assigned (Sorokin & Forsyth, 2008, p. 3). This might reflect a lower motivation and might in turn substantially affect research quality.

The problem of the opaque research process tends to be exacerbated by working with an anonymous group while at the same time opening up new possibilities for illuminating the problem. In the current study, we take advantage of the possibility to recruit a heterogeneous pool of coders through crowdworking and examine the effects on coding quality. In the next section we first elaborate on the central quality criteria of content analysis: Reliability and validity. We then discuss the factors that affect coding quality in general before we dive into best practices in using crowdworkers as coders. Through a comparative study on the discussion of abortion in German twitter, we not only systematically compare the coding quality of crowdworkers and student coders but also the influence of manifest and latent category types which differ in coding difficulty as well as coder characteristics in terms of personality traits and sociodemographics.

2. Reliability and validity

No matter what new developments emerge – whether or not work is done with automation or with crowdworkers – research practices must be judged by the quality criteria of any empirical research: Reliability and validity.

Reliability means the robustness of the measurement instrument: A measurement instrument – in content analysis, the codebook, or the category system therein – should lead to the same results when repeatedly applied to the same material (Hayes & Krippendorff, 2007; Krippendorff, 2016; Rössler, 2017, p. 205).

Intracoder reliability or stability refers to the extent to which the process remains the same over time (Brosius et al., 2022. p. 167; Kolb, 2004, p. 337; Krippendorff, 2013, p. 270). Stability is considered the weakest form of reliability and an insufficient criterion for classifying data as reliable since intracoder reliability cannot detect individual misinterpretations (Krippendorff, 2013, p. 270).

Intercoder reliability or replicability, in contrast, is a more stringent reliability measure. It aims at the agreement between several coders (Kolb, 2004, p. 337). Since most content analyses involve multiple coders, intercoder reliability is often especially important (Vogelgesang & Scharkow, 2012). Finally, researcher-coder reliability measures how closely the coding results match those of the researchers

(Rössler & Geise, 2013, p. 283). This type of reliability refers to the extent to which coding matches specifications and measures what it is supposed to measure (Krippendorff, 2013, p. 271). According to Krippendorff (2013, p. 271), researcher-coder reliability is the most stringent reliability measure because deviations also include deviations in stability and replicability. Thus, researcher-encoder reliability can also be considered part of reliability and equally as a criterion of validity (Früh, 2017, p. 189; Vogelgesang & Scharkow, 2012). To measure accuracy, test-standard procedures are used, i.e., the data produced by the coders are compared with data collected by other means, such as by the experts. If these external measures match with the coding, it is assumed to be accurate in the sense of the study (Krippendorff, 2013, p. 271; Song et al., 2020). The data coded by the experts can thereby be described as the gold standard.

The concepts of reliability and validity are interrelated (Potter & Levine-Donnerstein, 1999): reliability is a prerequisite for validity, a necessary but not sufficient condition (Krippendorff, 2013, p. 328; Lacy et al., 2015, p. 796; Potter & Levine-Donnerstein, 1999, p. 272). This is because reliability does not automatically ensure validity – data can be perfectly reliable yet invalid; but if a study is not reliable, the data can never be valid (Früh, 2017, p. 114; Hallgreen, 2012, pp. 24–25; Lombard et al., 2002, p. 589).

3. Influences on the coding behavior and results

Unboxing the coding process helps to improve our understanding of how coding decisions affect validity and reliability. Information processing models can provide important insights regarding the coding process (Wirth, 2001; Wirth et al., 2015). As a theoretical basis for understanding encoding processes, cognitive processes through which individuals transform incoming information into meaningful representations within their memory systems (Wirth et al., 2015), dual-process models are helpful. Dual-process models differentiate two distinct information processing systems operating simultaneously or sequentially: one characterized by deliberate, effortful, and conscious processing (central route), and another characterized by automatic, heuristic, and intuitive processing (peripheral route) – dependent on cognitive and motivational conditions (Chaiken, 1980; Petty & Cacioppo, 1986; Wirth, 2001; Wirth et al., 2015). In the case of the peripheral route (heuristic information processing), not all information in the codebook and the coding material is taken into account during coding, but certain cues in the text are sufficient to elicit a certain decision in coders. Such cues activate schemata in coders that act as heuristics to elicit a particular coding decision (Wirth et al., 2015, p. 99). In the case of the central route (reflective information processing), all information in the codebook and the coded material is considered, and no other influences are at work (Wirth et al., 2015, pp. 98–99). However, coders cannot always be assumed to be highly motivated and, therefore, reflective and systematic (Wirth, 2001, p. 168). To make matters worse, they are usually confronted with coding as much content as possible in as short as possible, and coding is a highly repetitive task.

The occurrence of the coding process, whether it follows the peripheral or central route, and the ultimate outcomes rely fundamentally on three factors: first, the *characteristics of the coding process* (chapter 3.1); second, *coder characteristics* (chapter 3.2); and third, the *category types* (chapter 3.3) that determines the type of variables to be coded.

3.1 Characteristics of the coding process

The most important characteristic of the coding process concerns the *material* being analyzed. Suppose the material to be coded is complex. In that case, coders might simplify the coding process, e.g., by using only a limited selection of category expressions or being unable to identify argumentative links in a complex text (Meyers & Brashers, 2010).

Equally relevant to the material is the design of the *codebook*. Bos (1989) shows that more detailed category descriptions improve intercoder reliability. In contrast, Spiegelman et al. (1953) could not provide evidence of such an effect: In their study, more detailed category information did not increase coding reliability – especially when coding more complex and interpretation-free content. This could be explained by the fact that more detailed category descriptions lead to coders using abbreviation strategies, processing them heuristically, and thus possibly no longer paying proper attention to the category descriptions (Degen, 2015, p. 88; Wirth, 2001, pp. 175–176). A reverse strategy of making codebooks short and simple is also likely to favor heuristic strategies because coders may be overly influenced by their characteristics or the characteristics of the coded material (Wirth, 2001, p. 176).

Finally, *coder training* plays a major role. Although the coding instructions should be available so that they can be understood by the coders in their written form (Neuendorf, 2017, p. 157) – it is common to provide coders with additional training. This involves teaching them how to reliably use the codebook and its coding rules so that coders with different backgrounds can also code consistently (Brosius et al., 2022, p. 168; Krippendorff, 2013; Neuendorf, 2017, p. 130). Nonetheless, there is a risk that coder training sessions establish implicit rules that coders and researchers share, which are not explicit and written down in the codebook (Krippendorff, 2013, p. 131).

3.2 Coder characteristics

Coders should be interchangeable, meaning that regardless of who codes the material in a study, the results should not vary (Neuendorf, 2017, p. 157). Notwithstanding, coders are subject to certain influencing factors that may affect the data.

Starting with *sociodemographic differences*, for example, Axelrod and Hone (2006) show that gender impacts the coding of emotional facial expressions. Coders of the same gender are more likely to agree with their assessments. Furthermore, the agreement is higher when the facial expression of a person of the opposite sex is to be estimated (Axelrod & Hone, 2006). Similar interactions and systematic biases are also reported by Al Kuwatly et al. (2020), who find that variance in gender, age, and native language lead to systematic biases in the annotation of hate speech.

Personal traits are another important factor to consider in the coding process. In general, it is preferable to choose coders who exhibit high levels of accuracy and attention to detail. Individuals with a tendency towards perfectionism may struggle when making decisions in ambiguous situations (Mallinger, 2018, p.103). Individuals with high levels of empathy or emotional sensitivity have been shown to be better at detecting and recalling emotional tendencies in text (Bloise & Johnson, 2006, p. 201). Traits such as conscientiousness or openness to experience may influence factors such as consistency and creativity.

Another coder characteristic is *coding experience*. Coders need increasingly less time to code an item as the number of coded items increases, i.e., they become more efficient (Wettstein et al., 2012; Wirth et al., 2015). This effect is evident for the first 20 to 50 coded units, but after that, the time required for coding hardly reduces (Wettstein et al., 2012). In the context of the coding experience within a specific content analysis, coders may be subject to learning and habituation effects: they develop a certain procedure and a certain way of coding the research material (Brosius et al., 2022, p. 176).

The *prior knowledge* of the coders can also influence the coding results. In some respects, a certain amount of prior knowledge is integral to successful coding: Only human coders can correctly interpret multiple modes of representation through their knowledge (Rössler, 2017, pp. 23–24). At the same time, the ideal content analysis requires that the prior knowledge of the coders is controlled by rules defined in the codebook so that uniform coding results are possible (Wirth, 2001, p. 172). According to Degen (2015, p. 84), it is also conceivable that differences in knowledge between coders will even out during the coding process since the knowledge of the respective topic area will likely increase as the number of coded contents increases.

Coders interested in the project in question or consider it particularly relevant might be more *motivated* to code correctly and accurately and, therefore, more likely to be reflective (Wirth et al., 2015, p. 101). Indeed, motivated coders do code more articles, but there was no such influence on the quality of coding (Wirth et al., 2015, p. 101). However, coders can be assumed to have low interest, which leads to lower motivation. Kronewald (2015) also examined the coder motivation and interviewed coders at a media analysis institute. Instead, a high level of intrinsic and extrinsic motivation was found. Some of the factors that have been identified as promoting higher motivation include payment, the working atmosphere, the opportunity to gain knowledge through coding, and the ability to effectively carry out tasks.

The type of information processing is also related to the *need for cognition*, which is a personality-specific factor (Petty & Cacioppo, 1986). Coders with a strong need for cognition may be more likely to code reflectively than coders with a low need for cognition (Wirth et al., 2015, p. 103). In this respect, it has been shown that coders with a high need for cognition code more reliably variables that are more difficult to code; this correlation was not found for simple variables (Wirth et al., 2015).

3.3 Category types

Given the presented characteristics of the coding process and coder characteristics, it is worth highlighting one more aspect. The manifest content is easy to observe and can be interpreted in the same way by all coders (Früh, 2017, p. 113). The epistemological interest of the social sciences is often not only in this manifest content but equally in the meanings hidden behind such manifest content, i.e., latent (Potter & Levine-Donnerstein, 1999). Measuring this latent content is more difficult than measuring manifest content. Potter and Levine-Donnerstein (1999, p. 261) distinguish three types of content to be coded: *manifest, latent*, and *projective*. In the context of coding, *projective* refers to a type of content that reflects the thoughts, feelings, or attributes projected by the person producing the content. Depending on the type of content to be coded, the coders' schemas play a greater role in coding, being more subordinate in the case of manifest content but more prominent in the case of projective content (Potter & Levine-Donnerstein, 1999, p. 262). Here it is already clear that the coding process is subject to certain influencing factors that emanate from the coders, among others.

Given manifest content, which requires few decisions from coders, poor agreement among coders indicates that they are not coding *correctly*. For example, they are assigning incorrect codes out of fatigue, which is why the data end up not being valid (Potter & Levine-Donnerstein, 1999, p. 271). When dealing with latent content, researchers need to design codebooks in such a way that the coding rules are so clearly stated that all coders make the same choices, and in this way, can achieve high reliability (Potter & Levine-Donnerstein, 1999, p. 272). Supposed the coding rules are designed in such detail as to ensure this. In that case, there is a risk that coders' attention will be drawn to aspects that can be coded more easily rather than capturing the aspects that should be measured. This can consequently limit the study's validity as the authors reflect (Potter & Levine-Donnerstein, 1999, p. 266). Also, if the categories of a codebook attempt to capture every aspect of a theoretical construct, this ensures high validity – but often at the expense of reliability (Brosius et al., 2022, p. 167). The same applies if the categories are too simple: While this may lead to coders coding reliably, the study's validity decreases (Krippendorff, 2013, p. 270). Even when coding rules are formulated with a degree of flexibility, providing coders with too much room for interpretation can result in potential risks. Consequently, in favor of higher validity, the reliability of the coding decreases (Früh, 2017, p. 120).

Holsti (1969) notes the identification of coding units as the primary task of coders, whereby it is necessary to separate the coding units from irrelevant content. He thereby differentiates according to the type of content. If a coding unit comprises a symbol or a section, fewer identification problems are expected than with thematic coding (Holsti, 1969, p. 136). While the boundaries of the coding unit are clearly indicated for the former (e.g., through specific symbols or indented paragraphs), such a physical description is lacking for thematic analyses where a single sentence may address multiple topics. This is problematic because delimitations can only be made through a judgment. A second task, after delimitation, is for coders to decide in which category the coding unit belongs (categorizing) and, if necessary, to decide in which subcategory it should be placed (subcategorizing) (Holsti, 1969, p. 137).

Potter and Levine-Donnerstein (1999, pp. 265–266), like Holsti (1969), distinguish between the type of content concerning coding tasks. They differentiate latent, manifest, and projective content. While they see the main task in coding manifest content as simply *typing* the occurrence, they describe the task for latent content as requiring patterns to be discovered. Once an indicator of the presence of a pattern to be coded is discovered in the material, a search is made for other indicators that speak to the presence of the pattern. If a sufficient number of indicators are found or if they are present in the required combination, this is recorded in the code sheet (Potter & Levine-Donnerstein, 1999, p. 265). Finally, with projective content, the task is mainly to make judgments based on one's own schemata (Potter & Levine-Donnerstein, 1999, pp. 265–266). The challenges for coders and, thus, the requirements for coder training are derived from this. For manifest content, coders must be enabled to recognize specific symbols. For latent content, they must also learn the rules that determine which symbol combinations fulfill the presence of a pattern (ibid., p. 266).

4. Crowdworkers as coders

When utilizing crowdworkers as coders, direct researcher-coder interaction is not possible. This makes the coding process even more opaque and in some ways uncontrollable. Coding with the crowd poses unique challenges, but potential advantages might outweigh these caveats and even foster new insights into the coding process. It is readily apparent that conducting content analysis with crowdworkers quickly and cost-effectively has a positive impact on the feasibility and flexibility of content analytic research (Guo et al., 2020, p. 812). A major advantage in regard to the question on how coding process affects research quality is the availability of a large and diverse pool of coders. Recruitment of coders in the crowd can result in more variation and a better representation of the population compared to convenience samples of students (Berinsky et al., 2012, p. 355). Paradoxically, while a diverse group of coders may help reduce bias in coding decisions and thus increase validity, it may also negatively affect the reliability of the measurement, as the heterogeneity may foster differences in coder judgments.

As highlighted in the previous section, coder heterogeneity might affect coding outcomes in numerous ways. Studies employing crowdworkers provide the opportunity to proactively address the issue. First and foremost, platforms maintain possibilities to assess crowdworkers' quality. Crowdworkers are classified into different levels, for example, on a trust score, the experience level, or the successfully completed tasks and approval rate. Second, crowd coding allows for coding tasks to be performed multiple times due to cost-effectiveness. This makes it possible to either identify underperformers and remove their data from the analysis or rely on majority decisions in the selection of final codings. It has already been shown that majority votes improve coding results (Budak et al., 2021, p. 149) and produce comparable results to the weighting of coding based on the trust score or an additional confidence rating (Guo et al. 2020, p. 825). However, such methodological safeguards reduce cost efficiency (Guo et al., 2020, p. 825). Third, in order to improve the motivation of coders and the quality of the coding, training tasks can precede the coding as well as tests that elicit knowledge and attitudes of potential coders (Budak et al., 2016; Budak et al., 2021, p. 146–153). Interspersed test questions can provide direct feedback (Lind et al., 2017, p. 197). Approaches known as "games with a purpose" (e.g., Prestopnik et al., 2017) go even further. They can increase the motivation of the crowdworkers and thus the data quality by using playful elements. Finally, ex-post tests, e.g., including the required time or compliance with a pre-coded sample, may also be used for quality assurance.

When decent quality assurance techniques are implemented, crowdworkers perform similarly well as student coders. In a comparative study, van Atteveldt et al. (2021) find, that humans outperform computational approaches in sentiment coding and that both, students as well as crowdworkers results were adaptable. Albeit, students provided higher compliance with a gold standard generated by the researchers indicating higher validity. In regard to reliability and resulting from a comparison of students vs. crowdworkers in sentiment analysis Lind et al. (2017) report similar results. Given the cost efficiency and speed, they conclude that "smaller deviations in ICR [inter coder reliability] may be a reasonable price to pay" (p. 204). However, the reported Krippendorffs alpha values are lower than the usual standard, especially for crowdworkers. In particular, the coding of latent constructs such as sentiment, which require a high level of interpretive performance, remain a challenge.

5. Research questions

The theoretical considerations show that various determinants influence the quality of codings. Firstly, coder characteristics seem relevant. More precisely, in this study, we want to investigate how sociodemographics such as age, gender, and education, as well as personality traits such as the need for cognition, and emotional sensitivity influence the coding quality. Secondly, the influence of easily codable manifest category types and challenging codable latent category types appears to be relevant. Lastly, the coding quality might also depend on whether the coder is a crowdworker or a student coder. We comprehend the coding quality by considering both its reliability and validity, as they serve as indicators of its overall effectiveness. We, therefore, ask the following research questions:

> *RQ1: To what extent does the coding quality (validity and reliability) differ between crowdworkers and student coders?*
>
> *RQ2: To what extent do latent and manifest category types differ in terms of coding quality (validity and reliability)?*
>
> *RQ3: To what extent are sociodemographics and personality traits as coder characteristics associated with the coding quality (validity and reliability)?*

6. Method

To answer the research questions, we employed a mixed-method design combining content analysis and survey. For the content analysis, we picked a task common in communication science and applied media research: the analysis of tweets. To keep this factor influencing coding quality constant in the study, we deliberately selected material that was short and supposedly easy to understand. The example topic focused on the current debate about legalizing abortion in Germany. In addition to its typicality, relevance, and topicality, this topic has further merits. It is controversial, so different opinions prevail in society, and it is complex to elicit manifest and latent variables.

The comparative design involved conducting the same content analysis twice: Four students served as coders in the first condition. In the second condition, the same task was performed using 150 crowdworkers as coders. All coders completed an online survey capturing their characteristics before coding. The aim is to determine the effects on the central quality measures of the content analysis as described above. These are reliability (comparison of coding between coders) and validity (comparison of coding between researchers and coders, the gold standard).

6.1 Sample of tweets

The sample consists of 300 German tweets that all contain at least one of the search terms *pro-choice*, *pro-life* or *abortion* (*Abtreibun*g in German) and were downloaded via the Twitter API in autumn 2019. The tweets could include the English terms pro-choice and pro-life, as these terms are also used in English in Germany. An initial screening ensured that all tweets referred to the topic and that the sample contained an equal proportion of pro-abortion, anti-abortion, and neutral/ambivalent tweets. The data was collected exclusively for the methodological study presented here.

6.2 Codebook and measures of category types

To address the research questions, a short codebook was developed. Contentwise, it addresses the issue of abortion, the positions taken in the Twitter debate, and the effects expressed by the users. The category types were chosen carefully to differ in terms of latent and manifest category types, i.e. to create different levels of difficulty. The first variable within the *latent category type* was (1) *positioning* of the writer. It contained the codes *pro-abortion* and *anti-abortion* and a remainder category for neutral or ambivalent tweets. As second variable within the *latent category type*, we captured (2) the *emotional valence* of the tweets. Coders were asked whether the statement was primarily factual or primarily emotional. It did not matter what emotion was expressed; the only issue was the valence. As a first variable for the manifest category type, coders had to decide whether or not there was a (3) *reference to the current legal situation* of abortion in Germany and the legal paragraphs under discussion. Finally, we included two formal variables as *manifest categories* that require concentration but left no room for interpretation:

coders were asked to count the number of (4) *question marks* and (5) *exclamation marks*.

Two researchers initially coded all tweets to generate a gold standard. This gold standard served as a reference for the validity of the coding of students and crowdworkers.

Although the study's top priority was not to achieve high validity and reliability – after all, some variance is necessary to answer the research questions – the instrument was still intended to emulate a realistic tool for capturing a content-related question. Hence, the codebook was pretested with additional student coders and optimized in terms of wording and clarity of instructions.

6.3 Measures of coder characteristics

To answer the research question regarding coder characteristics, we captured coders' *sociodemographic characteristics*, namely age, gender, and education, as well as the traits of *need for cognition* and *emotional sensitivity* on a Likert-like scale ranging from 1 = "strongly disagree" to 7 = "fully agree". The *need for cognition* was measured using an established short scale (Bless et al., 1994). Due to insufficient internal consistency, only three items were included in the final mean index (M = 4.59, SD = 1.09, ω_h = .64; N = 154). *Emotional sensitivity* was captured by combining approaches from Früh and Wünsch (2009) and Grimm (2015). All five items were compressed to a mean index (M = 5.30, SD = 0.917, ω_h = 0.804; N = 154). For a detailed overview of all items, see table 6 and 7 in the appendix.

6.4 Procedure

To integrate the online survey and the content analysis, we programmed online questionnaires using an online survey tool. To divide the task into a manageable size for an individual crowdworker, a total of 10 questionnaires were programmed, each containing 30 tweets in random order. Each tweet received its own questionnaire page on which the tweet, the short codebook, and an input mask were displayed. The input mask was programmed so that coders received feedback via a funny meme every five tweets on one of their past five codings. The feedback was based on the gold standard. This small gamification element served to increase the coders' motivation and, through that, to increase the quality of the coding. Before the coders started coding, they received detailed instructions and a short survey capturing their characteristics. The procedure was the same for student coders and crowd workers.

6.5 Comparative implementation in two coding conditions

Our research design included conducting the same study twice. In the first condition, four students who majored in communication science coded the whole sample of 300 tweets. To reassemble the typical conditions, the student coders were trained for one hour in the use of the codebook, while the crowdworkers were not trained – reflecting common research practice. Student coders were 21 to 24 years old (M = 22.5), and half were female. In the second condition, 150 crowdworkers were recruited via the platform Clickworker. They coded 30 tweets each without training. On average, coders were *M* = 40.3 years old (*SD* = 11.9), 38 percent were female, and 60 percent had a high education (German Abitur). In the crowdworker condition, each tweet was coded 15 times, while in the student condition, each tweet was coded four times. In total, the study includes 28,500 coding decisions. The provider Clickworker was chosen because it provided high-quality support and offered enough German-speaking coders. The data was collected via Clickworker within six hours. Since we wanted to offer the coders fair payment, they received 4.50 euros for the calculated 20 minutes of working time, in line with the minimum wage in Germany.

6.6 Data analysis

To answer our research questions on the influence of coder characteristics and category types on coding quality, we used two different measures of coding quality. First, we examined the influence of compliance with the gold standard as a measure of validity. Second, we investigated compliance with other coders as a measure of reliability. We first used descriptive analyses and compared the validity and reliability of crowdworkers and student coders to answer the first research question. We did not exclusively use Krippendorff, as the coefficient could not be calculated on a case level. To answer the research questions on the influence of category types (RQ2) and coder characteristics (RQ3), we used additional multilevel regression analyses. For the regression analysis with the agreement with the gold standard (validity) as the dependent variable, we calculated whether or not a coding decision was consistent with the gold standard as an indicator for validity (*M* = 0.73, *SD* = 0.44). Further, to address reliability, we calculated to what percentage a coding decision matched with fellow coders' decisions. The lowest value of 0 indicates that the coding does not match any other coder's coding. A value of 1, by contrast, indicates that the coding matches every other coder's coding. The resulting variable had a mean of *M* = 0.78 (*SD* = 0.29) and was considerably skewed (skew = -1.26, kurtosis = 0.37).

7. Results

7.1 Differences between crowdworkers and student coders (RQ1)

First, RQ1 regarding the difference between crowdworkers and student coders was answered with descriptive analyses. Regarding validity, the overall agreement with the gold standard was quite high – in most cases, the agreement between coders and researchers was over 80 percent (Table 1). The agreement was very low only in the latent category type of *emotional valence*, with 18 percent for student coders and 22 percent for crowdworkers. Regarding the differences between crowdworkers and student coders, significant differences were found in *positioning*, *emotional valence*, and *references to the law*. While the agreement with the gold standard was significantly higher for student coders for *positioning* and *reference to the law*, the agreement in *emotional valence* was significantly higher for crowdworkers – although it remained very low overall.

Table 1. Descriptive findings of compliance with the gold standard

	Student coders (*n* = 4 × 300 = 1,200)				Crowdworkers (*n* = 150 × 30 = 4,500)			
	M	*SD*	*SE*	*CI*	*M*	*SD*	*SE*	*CI*
Positioning*	0.83	0.38	0.01	[0.80, 0.85]	0.74	0.44	0.01	[0.73, 0.75]
Emotional valence*	0.18	0.39	0.01	[0.16, 0.21]	0.22	0.42	0.01	[0.21, 0.23]
Reference to law*	0.91	0.29	0.01	[0.89, 0.93]	0.86	0.35	0.01	[0.85, 0.87]
Question marks	0.94	0.25	0.01	[0.92, 0.95]	0.93	0.26	0.00	[0.92, 0.93]
Exclamation marks	0.91	0.29	0.01	[0.89, 0.92]	0.89	0.32	0.00	[0.88, 0.90]

Notes. The mean values indicate the level of compliance with the pre-coded gold standard ranging from 0 (0% compliance) and 1 (100% compliance). *95-percent confidence intervals do not overlap.

In terms of reliability, both agreement (Table 2) and Krippendorff (Table 3) were quite high for *question marks* and *exclamation marks*. Regarding the *coding of law*, the agreement was similarly high for both crowdworkers and student coders, but Krippendorff was much lower for crowdworkers. In contrast, *positioning* and *emotional valence* were relatively low for agreement and Krippendorff in crowdworkers and student coders. Overall, the reliability of agreement and Krippendorff was almost always significantly higher for students than crowdworkers.

Table 2. Descriptive findings of compliance with other coders (percent agreement)

	Student coders (*n* = 4 × 300 = 1,200)				Crowdworkers (*n* = 150 × 30 = 4,500)			
	M	*SD*	*SE*	*CI*	*M*	*SD*	*SE*	*CI*
Positioning*	.73	.24	0.01	[.72, .75]	.67	.30	0.00	[.66, .68]
Emotional valence*	.52	.24	0.01	[.50, .53]	.49	.24	0.00	[.49, .50]
Reference to law*	.91	.11	0.00	[.90, .91]	.84	.26	0.00	[.83, .85]
Question marks*	.97	.10	0.00	[.97, .98]	.96	.15	0.00	[.95, .96]
Exclamation marks	.90	.20	0.01	[.89, .91]	.88	.23	0.00	[.88, .89]

Notes. The mean values indicate the level of agreement from 0 (0% agreement) and 1 (100% agreement). *95-percent confidence intervals do not overlap.

Table 3. Student coders vs. crowdworkers: Comparison of Krippendorff's α

	Student coders (n = 4 × 300 = 1,200)			Crowdworkers (n = 150 × 30 = 4,500)		
	α	SE	CI	**α**	SE	CI
Positioning*	0.84	0.02	[0.80–0.88]	0.61	0.02	[0.56–0.65]
Emotional valence*	0.50	0.03	[0.44–0.56]	0.38	0.02	[0.34–0.42]
Reference to law*	0.96	0.01	[0.93–0.98]	0.62	0.02	[0.57–0.66]
Question marks*	0.93	0.02	[0.89–0.97]	0.84	0.02	[0.80–0.87]
Exclamation marks*	0.87	0.02	[0.82–0.91]	0.75	0.02	[0.70–0.79]

Notes. *95-percent confidence intervals do not overlap. 1000 Bootstrap samples, drawn from coding units.

7.2 The influence of coder characteristics (RQ2) and category type (RQ3)

Multilevel logistic regressions were used to answer research questions regarding the association between coding quality and coder characteristics (RQ2) and between coding quality and category type (RQ3). Regarding validity, the first multilevel logistic regression with compliance with the gold standard as the dependent variable showed that compliance was mainly influenced by category types (Table 4). Model 1 shows the null model. Although the ICC value for the coders is very small (0.02), the value ICC = 0.04 of the tweets suggests to keep the multi-level structure of the data. In light of the manifest and latent category types, Model 2 reveals that the reference category *positioning,* used due to dummy coding, significantly differs from all other categories regarding the association with the gold standard. Specifically, *emotion valence* deviates in a negative direction from *positioning*, while *reference to law*, *question marks*, and *exclamation marks* deviate positively. As Model 2 reveals, the reference categories of all latent and manifest variables of category types significantly differed associated with compliance with the gold standard. *Emotional valence* as latent category type has a negative influence on compliance with the gold standard, while the manifest categories *reference to law* and number of *punctuation marks* are positively associated with this measure of validity. The explanatory power of the model is quite high with pseudo R^2 = .42.

In contrast, Model 3, which includes the coder characteristics, does not provide any additional explanatory power. Relatively, the regression weights of the coder characteristics are very small, and the weights of the coefficients outweigh their effects. The significant *b*-values of *emotional sensitivity* and *education* can nonetheless be interpreted as hints that these coder characteristics may influence the coding result more than others.

Table 4. Multilevel logistic regression result for compliance with the gold standard

	Model 1 *b (SE)*	Model 2 *b (SE)*	Model 3 *b (SE)*
Fixed effects			
Intercept	0.73*** (0.01)	0.75***(0.01)	0.75***(0.01)
Category type			
Emotional valence		-0.54***(0.01)	-0.54***(0.01)
Reference to law		0.11***(0.01)	0.11***(0.01)
Question marks		0.17***(0.01)	0.17***(0.01)
Exclamation marks		0.13***(0.01)	0.13***(0.01)
Coder characteristics			
Need for cognition			-0.00 (0.00)
Emotional sensitivity			0.01***(0.00)
Age			0.01(0.01)
Gender (1 = male, 2 = female)			0.01(0.00)
Formal education			0.02***(0.00)
Random effects			
Standard deviation tweet	0.09	0.10	0.10
Standard deviation coder	0.06	0.06	0.05
Goodness of fit			
AIC	33,459	19,840	19,769
$\Delta\chi2$		13,662	38
$\Delta \mathit{df}$		4	5
Pseudo-R^2 (total)	0.06	0.42	0.42

Note. *N* = 28,500 coding decisions; Coder Characteristics were scaled and grand mean centered; standard errors are in parentheses. ICC of Model 1: Tweet = 0.04; Coder = 0.02. All *p* values are two-tailed.

*** *p* < .001; ** *p* < 01; * *p* < 05

Next, we present the equivalent analysis for reliability, and a similar picture emerges. Again, the ICC values argue for the analysis in the multilevel structure (ICC_{Tweet} = 0.05; ICC_{Coder} = 0,06). The dependent variable here is compliance with codings of the same coding unit by other coders, ranging from 0 = *no other coder coded the same value on a particular tweet* to 1 = *all coders coded the same value on the* *particular tweet*. Since the dependent variable represents fractional logits and is highly skewed (-1.26), Table 5 shows ordered beta regressions (Kubinec, 2022). Like validity, reliability is also mainly influenced by category types and not so much by coders' characteristics. Again, in contrast to reference category positioning, all category types significantly influenced compliance with others (Model 2). Emotional sensitivity, age, and education were significant for coders' characteristics but reveal only minimal effects (Model 3).

Table 5. Multilevel ordered beta regression result for compliance with other coders

	Model 1 *b (SE)*	Model 2 *b (SE)*	Model 3 *b (SE)*
Fixed effects			
Intercept	**0.71*** (0.03)**	**0.50***(0.03)**	**-0.42***(0.18)**
Category type			
Emotional valence		**-0.61***(0.02)**	**-0.61***(0.02)**
Reference to law		**0.64***(0.02)**	**0.64***(0.02)**
Question marks		**1.73***(0.03)**	**1.73***(0.03)**
Exclamation marks		**1.04***(0.02)**	**1.04***(0.02)**
Coder characteristics			
Need for cognition			-0.01(0.02)
Emotional sensitivity			**0.06*(0.03)**
Age			**0.00*(0.00)**
Gender (1 = male, 2 = female)			-0.01(0.05)
Formal education			0.12***(0.03)
Random effects			
Standard deviation tweet	0.32	0.11	0.02
Standard deviation coder	0.22	0.28	0.06
Goodness of fit			
AIC	32,947	22,042	22,020
Δχ2		10,914	31
Δ*df*		4	5

Note. N = 28,481 coding decisions; standard errors are in parentheses. All *p* values are two-tailed. *** *p* < .001; ** *p* < 01; * *p* < 05.

8. Conclusion

This study aimed to examine to what extent coders characteristics and manifest and latent category types influence the coding quality and to what extent the coding quality differs between crowdworkers and student coders. The coding quality was represented by the agreement with the gold standard (validity) and the percentual agreement within coders and Krippendorff's alpha (reliability).

Regarding the difference in coding quality between crowdworkers and student coders (RQ1), student coders showed a significantly better validity for most category types. In terms of reliability, there were mixed results. On the one hand, student coders received a higher quality for *positioning* and *reference to the law*. On the other hand, crowdworkers were significantly better in *emotional valence* – however, the *emotional valence* was still quite low. The fact that students often coded in better quality could be due to the coding training that students received, but crowdworkers did not. Students may have developed a better understanding of coding by reviewing and discussing different examples and by a similar understanding as a result of the shared training. Although our approach reflects current research practices, prior training can also improve reliability in crowdcoding (Budack et al., 2021). The higher coding quality of students could also be due to education, as some of the crowdworkers have lower education. This assumption is supported by the logistic regressions, which show that education significantly impacts coding quality. In addition, students might also have more experience with the content analysis method and therefore be more proficient. The differences in coding quality between crowdworkers and student coders implies that students are better suited for high quality in manual content analysis. Obvious reasons are routine and comprehension of the method, better control of the research process, and direct interaction between researcher and coder. That being said, especially for simple coding tasks, the quality of crowdworkers can definitely keep up with that of students. Hence, researchers can use crowdworkers here without major concerns. Especially if researchers are looking for an alternative due to time constraints or large amounts of data, crowdworkers represent a reasonably good alternative. Overall, the study thus follows Lind and colleagues (2017), who see crowdworkers as a reliable and valid alternative to students. This is especially true when it is recognized in research practice that quality assurance research practices like coder training are also necessary in crowdcoding (Budak et al., 2021). Nevertheless, right now students might still be the better choice when tasks are more difficult. However, whether with crowdsourced or with student coders, our findings provide guidance on the criteria that should be used to select coders.

The results of the regression analysis also demonstrate that it is the category types in particular that determine the coding result (RQ2). All category types significantly differed in their impact on coding quality when compared to the reference category *positioning*. The latent variable *emotional valence* differed particularly strong from *positioning* by negatively influencing the coding quality. From the descriptive results, we know that the coders of both groups were not able to code this category validly or consistently. Possibly coders did too much interpretation, resulting in a different perception of *emotional valence* between coders (reliability) and between coders and the gold standard (validity). Further, the fact that coders perform well on manifest categories such as counting tasks while being insufficient on latent categories such as semantic interpretation holds important implications with respect to the automation of content analysis. Mere counting and coding of manifest constructs can be automated even using simple computational techniques such as dictionaries. For semantic interpretations, human coders are urgently needed. But the fact that human coders are hardly able to produce valid and reliable results, and only with a great deal of effort, has implications for manual analysis of media content but also for the automation of this process. The risk arises that humans' poor codings of latent constructs will be adopted and replicated by automated procedures.

Regarding the influence of coder characteristics (RQ3), the factors tested in this study had minor influence on coding quality. Only *age*, *education* and *emotional sensitivity* were significantly associated with a higher validity and reliability in codings. *Age* and *education* as influencing factors suggest that experience, basic knowledge as well as intellectual competencies enhance coding quality. Yet, it is not the joy of thinking or the fun of tricky tasks that leads to the desired result here, as *need for cognition* is not an influencing factor. Admittedly, coding is probably a rather boring task from the coder's point of view. Therefore, it is not advisable to solely focus on the cognitive skills of coders. As demonstrated by the influence of *emotional sensitivity*, other personality traits can contribute to improving coding quality as well. Emotionally sensitive individuals may have a better ability to empathize with texts and are more attentive to details, subtexts, or latent meanings. Additionally, they may generally be more accurate in their work or have a stronger motivation to meet the researchers' demands.

Like any empirical work, this study has limitations that restrict its conclusiveness and may guide future research. In addition to the overall low validity, which is shown by a lack of agreement with the gold standard, it should be mentioned here in particular that the reliability values of the two groups are not well comparable. Specially for Krippendorff's comparison, it is problematic that in the case of crowdworkers, many coders coded only a few tweets each, and in the case of student coders, only a few coders coded many tweets. Future studies should look for a more balanced approach. Another imbalance regards the fact that student coders received training while crowdworkers did not. Although this approach reflects common research practice, it may account for some difference between the two groups, since it is advisable to provide crowdworkers with an adjusted training to increase coding quality (see Budak et al., 2021). Furthermore, it can be questioned whether the gold standard is the ideal measure of criterion validity. In this study, it was used as a suitable measure of internal validity, but with regard to external validity, a researcher generated gold standard would be too short-sighted, as the measurement depends on the training and the approach of the researchers (Song et al., 2020; Völker & Scholl, 2016). Further, the gold standard produced by the researchers and the codings of the student coders may show similar results because they have a more similar privileged position from which to look at the media content. Both groups have high educational backgrounds. In terms of external validity it may even be desirable to gain a broader view of the analysed media content through more heterogeneity among the coders. Beyond that, the coding characteristics we chose represent only a small selection of possible influencing variables. Other variables could also contribute to heterogeneous coding. For instance, in addition to other personality traits, attitudes, and prejudices are likely influencing factors (Früh, 2005). It should also be asked whether a homogeneous or a heterogeneous pool of coders is desirable. In addition to the coder characteristics, the process of coding itself can also be investigated more closely. In our study, gamification was used as a motivational tool, but its effectiveness was not assessed as it was not part of the experimental variation. Future studies could compare coders' motivation and coding with and without gamification elements. Furthermore, the theory section of this paper discussed other aspects that could influence the coding process like different features of the codebook or the coding situation.

This study has once again shown how difficult it is to produce valid results in standardized content analyses. Since this is merely a methodological exploration, the fragile reliability and validity values do not diminish the results – in a way, variation in the coding was necessary to reveal them. But methodological research aside, insufficient values are all too often accepted in quantitative content analysis and the goodness of coding is not reported to an adequate extent (Vogelgesang & Scharkow, 2012). In addition to a reflection on which constructs are measureable at all and should be included in the codebook, a comprehensive documentation of the reliability checks is necessary. This should also include indication of the error-prone nature of the measurement, for example by reporting standard error and confidence intervals.

Finally, the results of our study need to be reflected in the light of current developments in the field of automated content analysis. Computers have the merit of applying rules consistently and being more efficient than human coders, which is why automated analysis can reduce costs and process massive amounts of data (Lacy et al., 2015). Until recently, capturing the semantic meaning was a challenge in automated analysis and therefore the linguistic competence of coders was still necessary (van Attefeldt et al., 2021, Scharkow, 2013). However, large language models like GPT (OpenAI, 2023) have improved in this regard and future improvements are to be expected, even in difficult areas such as the coding of humor and irony. A comparison by Gilardi et al. (2023) shows that GPT-3.5 already achieves higher reliability than both, research assistants and crowdworkers. Further, it also outperformed the crowdworkers in compliance with the codes generated by research assistants. Although these results are remarkable, these developments are not a reason to speculate about the obsolescence of human coding in content analysis. Instead, we need to consider consequences for coding quality. Just like human coders, algorithms are subject to biases, which primarily stem from the training data used (Fazelpour & Danks, 2021). High-level tools like GPT can appear as an authority and seem detached from their training data. The formation of their responses is a black box in itself. It is challenging for individual researchers to effectively monitor and address biases. Human coding is still essential as an external validity criterion, so it is imperative to continue to examine the conditions of the human coding process. In addition, the fruitful combination of automated and human coding has already been discussed in the context of simpler automation approaches, such as the customization of dictionaries (Haselmayer & Jenny, 2017; Reveilhac & Morselli, 2022) and topic modeling (Baden et al., 2020). Because of its easy accessibility and the ability to argue for or against a particular coding, GPT has special potential in hybrid analysis. It could therefore play an important role in making the coding in quantitative content analysis more comprehensible, reliable and valid. How such a process will be optimally designed and what role the selection of (crowdsourced) coders will have in this process needs to be shown in further research.

Author note

Anonymized data as well as reproduction code and the full codebook are available at https://osf.io/9ug6s/.

Acknowledgments

We gratefully thank Deutsche Gesellschaft für Online Forschung (DGOF) for supporting this research with a research grant. Also, we thank the editors and two anonymous reviewers for their insightful feedback.

References

Al Kuwatly, H., Wich, M., & Groh, G. (2020, November). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 184–190). https://aclanthology.org/2020.alw-1.21.pdf

Axelrod, L., & Hone, K. S. (2006). Affectemes and allaffects. A novel approach to coding user emotional expression during interactive experiences. *Behaviour & Information Technology*, *25*(2), 159–173. https://doi.org/10.1080/ 01449290500331164

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures*, *14*(3), 165–183. https://doi.org/10.1080/19312458.2020.1803247

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295. https://doi.org/10.1017/S0003055416000058

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political analysis*, *20*(3), 351–368. https://doi:10.1093/pan/mpr057

Bos, W. (1989). Reliabilität und Validität in der Inhaltsanalyse. Ein Beispiel zur Kategorienoptimierung in der Analyse chinesischer Textbücher für den muttersprachlichen Unterricht von Auslandschinesen [Reliability and validity in content analysis. An example of category optimization in the analysis of Chinese textbooks for teaching Chinese abroad as a mother tongue]. In W. Bos & C. Tarnai (Eds.), *Angewandte Inhaltsanalyse in empirischer Pädagogik und Psychologie* (pp. 61–72). Waxmann.

Boxman-Shabtai L. (2021). Encoding polysemy in the news. *Journalism*, *24*(5), 1089–1108. https://doi:10.1177/14648849211045963

Bloise, S. M., & Johnson, M. K. (2007). Memory for emotional and neutral information: Gender and individual differences in emotional sensitivity. *Memory*, *15*(2), 192–204. https://doi.org/10.1080/09658210701204456

Brosius, H.-B., Haas, A., & Unkel, J. (2022). *Methoden der empirischen Kommunikationsforschung. Eine Einführung* [Methods of empirical communication Research. An introduction] (8th ed.). Springer VS. https://doi.org/10.1007/978-3-658-34195-4

Budak, C., Garrett, R. K., & Sude, D. (2021). Better crowdcoding: Strategies for promoting accuracy in crowdsourced content analysis. *Communication Methods and Measures*, *15*(2), 141–155. https://doi.org/10.1080/19312458.2021.1895977

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(1), 250–271. https://doi.org/10.1093/poq/nfw007

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Degen, M. (2015). Codierer-Effekte in Inhaltsanalysen: Ein vernachlässigtes Forschungsfeld [Coder effects in content analyses: A neglected field of research]. In W. Wirth, K. Sommer, M. Wettstein, & J. Matthes (Eds.), *Qualitätskriterien in der Inhaltsanalyse* (pp. 78–95). Halem.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, *16*(8), e12760. https://doi.org/10.1111/phc3.12760

Früh, W. (2005). Fortschritte bei der Inhaltsanalyse [Advances in content analysis]. In J. Wilke (Eds.), *Die Aktualität der Anfänge. 40 Jahre Publizistikwissenschaft an der Johannes-Gutenberg-Universität Mainz* (pp. 115–124). Halem.

Früh, W. (2017). *Inhaltsanalyse. Theorie und Praxis* [Content Analysis. Theory and practice] (8th ed.). UVK.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv*. https://doi.org/10.48550/arXiv.2303.15056

Guo, L., Mays, K., Lai, S., Jalal, M., Ishwar, P., & Betke, M. (2020). Accurate, fast, but not always cheap: Evaluating "crowdcoding" as an alternative approach to analyze social media data. *Journalism & Mass Communication Quarterly*, *97*(3), 811–834. https://doi.org/10.1177/1077699019891437

Hallgreen, K. A. (2012). Computing inter-rater reliability for observational data. An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.

Hase, V. (2022). Automated content analysis. In: F. Oehmer-Pedrazzi, S. H. Kessler, E. Humprecht, K. Sommer, & L. Castro (Eds.), Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft (pp. 23–36). Springer. https://link.springer.com/chapter/10.1007/978-3-658-36179-2_3

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity*, *51*(6), 2623–2646. https://doi.org/10.1007/s11135-016-0412-4

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley.

Hornik, R.; Binns, S., Emery, S., Maidel Epstein, V., Jeong, M., Kim, K., Kim, Y., Kranzler, E. C., Jesch, E., Juhyun Lee, S., Levin, A. V., Liu, J., O'Donnell, M. B., Siegel, L., Tran, H., Williams, S., Yang, Q., & Gibson, L. A. (2021). The effects of tobacco coverage in the public communication environment on young people's decisions to smoke combustible cigarettes. *Journal of Communication*, *72*(2), 187–213. https://doi.org/10.1093/joc/jqab052

Hornuf, L., & Vrankar, D. (2022). Hourly wages in crowdworking: A meta-analysis. *Business & Information Systems Engineering*, *64*(5), 553–573. https://doi.org/10.1007/s12599-022-00769-5

Kolb, S. (2004). Verlässlichkeit von Inhaltsanalysedaten. Reliabilitätstest, Errechnen und Interpretieren von Reliabilitätskoeffizienten für mehr als zwei Codierer [Reliability of content analysis data. Reliability test, calculate and interpret reliability coefficients for more than two coders]. *Medien & Kommunikationswissenschaft*, *52*(3), 335–354. https://doi.org/10.5771/ 1615-634x-2004-3-335

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research. An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research*, *18*(2), 243–250.

Krippendorff, K. (2013). *Content analysis. An introduction to its methodology* (3rd ed.). Sage.

Krippendorff, K. (2016). Misunderstanding reliability. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *12*(4), 139–144. https://doi.org/10.1027/1614-2241/a000119

Kronewald, E. (2015). Blackbox Feldphase: Strategien für die Motivation, Effektivität und Effizienz von Codierern [Black box field phase: Strategies for coder motivation, effectiveness, and efficiency]. In W. Wirth, K. Sommer, M. Wettstein, & J. Matthes (Eds.), *Qualitätskriterien in der Inhaltsanalyse* (pp. 140–158). Halem.

Kubinec, R. (2022). Ordered beta regression: A parsimonious, well-fitting model for continuous data with lower and upper bounds. *Political Analysis*, *31*(4), 519– 536. https://doi.org/10.1017/pan.2022.20

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*, *9*(4), 791–811. https://doi.org/10.1177/1077699015607338

Lauf, E. (2001). ».96 nach Holsti«. Zur Reliabilität von Inhaltsanalysen und deren Darstellung in kommunikationswissenschaftlichen Fachzeitschriften [».96 according to Holsti«. On the reliability of content analyses and their presentation in communication science journals]. *Publizistik*, *46*(1), 57–68. https://doi.org/10.1007/s11616-001-0004-7

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication Methods and Measures*, *11*(3), 191–209. https://doi.org/10.1080/19312458.2017.1317338

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication. Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Lovejoy, J., Watson, B. R., Lacy, S., & Riffe, D. (2014). Assessing the reporting of reliability in published content analyses 1985–2010. *Communication Methods and Measures*, *8*(3), 207–221. https://doi.org/10.1080/ 19312458.2014.937528

Lovejoy, J., Watson, B. R., Lacy, S., & Riffe, D. (2016). Three decades of reliability in communication content analyses. Reporting of reliability statistics and coefficient levels in three top journals. *Journalism & Mass Communication Quarterly*, *93*(4), 1135–1159. https://doi.org/10.1177/1077699016644558

Mallinger, A. (2009). The myth of perfection: Perfectionism in the obsessive personality. *American Journal of Psychotherapy*, *63*(2), 103–131. https://doi.org/10.1176/appi.psychotherapy.2009.63.2.103

Merten, K. (1995). *Inhaltsanalyse. Einführung in Theorie, Methode und Praxis* [Content analysis. Introduction to theory, method and practice] (2nd ed.). Westdeutscher Verlag.

Meyers, R. A., & Brashers, D. (2010). Extending the conversational argument coding scheme. Argument categories, units, and coding procedures. *Communication Methods and Measures*, *4*(1-2), 27–45. https://doi.org/10.1080/ 19312451003680467

Neuendorf, K. A. (2017). *The content analysis guidebook* (2nd ed.). Sage.

OpenAI (2023). *GPT-4 is OpenAI's most advanced system, producing safer and more useful responses.* https://openai.com/gpt-4

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*(3), 258–284. https://doi.org/10.1080/00909889909365539

Prestopnik N., Crowston K., & Wang J. (2017) Gamers, citizen scientists, and data: Exploring participant contributions in two games with a purpose. *Computers in Human Behavior 68*(2), 254–268. https://doi.org/10.1016/j.chb.2016.11.035

Reveilhac, M., & Morselli, D. (2022). Dictionary-based and machine learning classification approaches: A comparison for tonality and frame detection on Twitter. *Political Research Exchange*, *4*(1), 2029217. https://doi.org/10.1080/2474736X.2022.2029217

Riffe, D., Lacy, S., & Fico, F. G. (2005). *Analyzing media messages. Using quantitative content analysis in research* (2nd ed.). Routledge.

Rössler, P. (2017). *Inhaltsanalyse* [Content Analysis] (3rd ed.). UVK.

Rössler, P., & Geise, S. (2013). Standardisierte Inhaltsanalyse: Grundprinzipien, Einsatz und Anwendung [Standardized content analysis: Basic principles, use and application]. In W. Möhring & D. Schlütz (Eds.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (pp. 269–287). Springer.

Scharkow, M. (2013). Automatische Inhaltsanalyse [Automatic content analysis]. In W. Möhring & D. Schlütz (Eds.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (pp. 289–306). Springer. https://doi.org/10.1007/978-3-531-18776-1_16

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with amazon mechanical turk. *Computer vision and pattern recognition workshops, 2008. CVPRW'08. IEEE computer society conference on IEEE* (pp. 1–8).

Spiegelman, M., Terwilliger, C., & Fearing, F. (1953). The reliability of agreement in content analysis. *Journal of Social Psychology*, *37*(2), 175–187.

van Atteveldt, W., Van der Velden, M. A., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Völker, J., & Scholl, A. (2016, November 9–12). *Beyond statistical reasoning on reliability in standardized content analysis. A constructivist reflection about the social aspect of methodology* [Conference presentation]. 6th European Communication Conference of the European Communication Research and Education Association (ECREA), Prague, Czech Republic.

Vogelgesang, J., & Scharkow, M. (2012). Reliabilitätstests in Inhaltsanalysen. Eine Analyse der Dokumentationspraxis in Publizistik und Medien & Kommunikationswissenschaft [Reliability testing in content analysis. An analysis of documentation practice in Publizistik and Medien & Kommunikationswissenschaft]. *Publizistik*, *57*(3), 333–345. https://doi.org/10.1007/s11616-012-0154-9

Wettstein, M., Reichel, K., Kühne, R., & Wirth, W. (2012, April 20–21). *IN-TOUCH – ein neues Werkzeug zur Prüfung und Bewertung von Codiereraktivitäten bei der Inhaltsanalyse* [IN-TOUCH – a new tool for testing and evaluating coder activities in content analysis] [Conference presentation]. 13th Annual Meeting of the Swiss Society for Communication and Media Studies (SGKM), Neuchâtel, Switzerland.

Wirth, W. (2001). Der Codierprozeß als gelenkte Rezeption. Bausteine einer Theorie des Codierens [The coding process as guided reception. Building Blocks of a Theory of Coding]. In W. Wirth & E. Lauf (Eds.), *Inhaltsanalyse. Perspektiven, Probleme, Potentiale* (pp. 157–182). Halem.

Wirth, W., Wettstein, M., Kühne, R., & Reichel, K. (2015). Theorie und Empirie des Codierens: Personelle und situative Einflussfaktoren auf Qualität und Quantität des Codierens bei der Inhaltsanalyse [Theory and empirics of coding: Personal and situational factors influencing quality and quantity of coding in content analysis.]. In W. Wirth, K. Sommer, M. Wettstein, & J. Matthes (Eds.), *Qualitätskriterien in der Inhaltsanalyse* (pp. 96–118). Halem.

Appendix

Table 6. Item statistics: need for cognition

Item	*M*	*SD*
I like my life to be full of tricky problems to solve.	4.36	1.41
I would prefer more complicated problems to simple problems.	4.05	1.49
First and foremost, I think because I have to. (inverted)	5.35	1.58
Excluded: It is sufficient for me just to know an answer without understanding the reasons for the answer of a problem. (inverted)	5.10	1.58
Mean index (ω_h = .64; *N* = 154)	4.59	1.09

Table 7. Item statistics: emotional sensitivity

Item	*M*	*SD*
I can empathize with the emotional state of others very easily.	5.30	1.22
It touches me when I see a stranger or outsider isolated and lonely in a group.	5.41	1.19
When I see someone being insulted and humiliated, it touches me and I want to help.	5.45	1.27
Problems are rarely black or white – usually the truth lies somewhere in the middle.	4.89	1.24
I usually feel it‘s quite easy to see things from another person‘s point of view.	5.44	1.24
Mean index (ω_h = 0.804; *N* = 154)	5.30	0.92