SC|M
Studies in Communication and Media

FULL PAPER

Meaning multiplicity and valid disagreement in textual measurement: A plea for a revised notion of reliability

Mehrdeutigkeit und valide Nichtübereinstimmung in der textuellen Messung: Ein Plädoyer für einen revidierten Reliabilitätsbegriff

Christian Baden, Lillian Boxman-Shabtai, Keren Tenenboim-Weinblatt, Maximilian Overbeck & Tali Aharoni

Christian Baden (Prof.), The Hebrew University of Jerusalem, Department of Communication and Journalism, Mount Scopus, 9190501 Jerusalem, Israel. Contact: c.baden(at)mail.huji.ac.il. ORCID: https://orcid.org/0000-0002-3771-3413
Lillian Boxman-Shabtai (Dr.), The Hebrew University of Jerusalem, Department of Communication and Journalism, Mount Scopus, 9190501 Jerusalem, Israel. Contact: lilly.boxman(at)mail.huji.ac.il. ORCID: https://orcid.org/0000-0003-4893-6425
Keren Tenenboim-Weinblatt (Prof.), The Hebrew University of Jerusalem, Department of Communication and Journalism, Mount Scopus, 9190501 Jerusalem, Israel. Contact: keren.tw(at)mail.huji.ac.il. ORCID: https://orcid.org/0000-0001-9268-3969
Maximilian Overbeck (Dr.), The Hebrew University of Jerusalem, Department of Communication and Journalism, Mount Scopus, 9190501 Jerusalem, Israel. Contact: m.overbeck(at)mail.huji.ac.il. ORCID: https://orcid.org/0000-0003-3658-5584
Tali Aharoni (MA), The Hebrew University of Jerusalem, Department of Communication and Journalism, Mount Scopus, 9190501 Jerusalem, Israel. Contact: tali.aharoni(at)mail.huji.ac.il. ORCID: https://orcid.org/0000-0002-2138-8329

FULL PAPER

Meaning multiplicity and valid disagreement in textual measurement: A plea for a revised notion of reliability

Mehrdeutigkeit und valide Nichtübereinstimmung in der textuellen Messung: Ein Plädoyer für einen revidierten Reliabilitätsbegriff

Christian Baden, Lillian Boxman-Shabtai, Keren Tenenboim-Weinblatt, Maximilian Overbeck & Tali Aharoni

Abstract: In quantitative content analysis, conventional wisdom holds that reliability, operationalized as agreement, is a necessary precondition for validity. Underlying this view is the assumption that there is a definite, unique way to correctly classify any instance of a measured variable. In this intervention, we argue that there are textual ambiguities that cause disagreement in classification that is not measurement error, but reflects true properties of the classified text. We introduce a notion of *valid disagreement*, a form of replicable disagreement that must be distinguished from replication failures that threaten reliability. We distinguish three key forms of meaning multiplicity that result in valid disagreement – ambiguity due to under-specification, polysemy due to excessive information, and interchangeability of classification choices – that are widespread in textual analysis, yet defy treatment within the confines of the existing content-analytic toolbox. Discussing implications, we present strategies for addressing valid disagreement in content analysis.

Keywords: Content analysis, reliability, measurement validity, meaning multiplicity, ambiguity, polysemy.

Zusammenfassung: Reliabilität, operationalisiert als Übereinstimmung zwischen Codierern, ist gemeinhin akzeptiert als eine notwendige Voraussetzung für Validität in der quantitativen Inhaltsanalyse. Diese Sichtweise beruht auf der Annahme, dass für jedes Vorkommen einer gemessenen Variable exakt eine korrekte Klassifizierung bestimmt werden kann. Entgegen dieser Annahme argumentieren wir hier, dass divergierende Klassifizierungen auch aufgrund von im Text verankerten Mehrdeutigkeiten entstehen können, welche nicht als Messfehler zu verstehen sind, sondern vielmehr relevante Eigenschaften des analysierten Materials widerspiegeln. Entsprechend entwerfen wir einen Begriff *valider Nichtübereinstimmung* im Klassifizierungsprozess, welcher es erlaubt zwischen textuell verankerten, replizierbaren Mehrdeutigkeiten und Replikationsfehlern, welche die Reliabilität der Messung unterminieren, zu unterscheiden. Wir unterscheiden drei in der Textanalyse weit verbreitete Formen von Mehrdeutigkeiten – begründet in Unter-Spezifizierung, Informationsüberschuss, sowie der Austauschbarkeit von Klassifizierungsmöglichkeiten – welche valide

Nichtübereinstimmung zur Folge haben, aber bislang keine hinreichende Beachtung im inhaltsanalytischen Methodenkanon gefunden haben. Aufgrund einer Diskussion möglicher Implikationen dieser Auslassung skizzieren wir geeignete Strategien für den Umgang mit valider Nichtübereinstimmung in der inhaltsanalytischen Forschung.

Schlagwörter: Inhaltsanalyse, Reliabilität, Validität der Messung, Mehrdeutigkeit.

1. Introduction

In quantitative content analysis, conventional wisdom holds that reliability, operationalized as inter-coder agreement, is a necessary precondition for validity. Relying on identical coding rules, different coders must be able to classify the same coding units in consistent ways. In the words of the late Klaus Krippendorff, one of the sharpest minds invested into this debate: “We need to measure the extent of *agreement* among independent *replications* in order to estimate whether we can *trust* the generated data in subsequent analyses” (2016, p. 140; emphasis added). Of course, “reliability is only a prerequisite to validity” (Krippendorff, 2011, p. 94): Measurement can be replicable but non-valid.[1] However, if attempts at replication fail to produce agreement, measurement is considered unreliable, and no valid inferences can be made.

Underlying the convention of attributing all disagreement to error in the coding processes is the assumption that there is a definite, unique way to correctly classify any instance of a measured variable (e.g., Reidsma & Carletta, 2008). Accordingly, disagreement may either stem from coder bias – idiosyncratic variations in how different individuals interpret and apply the same coding rules; or from random error – unstructured mistakes in the recording of classification decisions (Krippendorff, 2008).[2] While there are some occasional nods in the literature to the challenges raised by textual ambiguity (e.g., Krippendorff, 2004) or the “difficulty” of coded categories (e.g., Zhao et al., 2022), the dominant view is that such challenges simply cause higher rates of coder bias and random error. The prevalent solutions are to either “force” agreement by further explicating and narrowing coding instructions (Craggs & Wood, 2005), or to dismiss failed variables as unreliable.

In this intervention, we take issue with this contention. Building upon extant scholarship in qualitative textual research (e.g., Ceccarelli, 2001; Eco, 1979; Joseph, 2018) and reception studies (e.g., Fiske, 1989; Hall, 1980), as well as our own long-standing experience in quantitative textual analysis, we argue that the inherent meaning multiplicity of some texts (Boxman-Shabtai, 2020) causes disagreement in classification that is not well-understood as measurement error, but reflects true properties of the classified text. After a brief review of current conceptualizations

[1] Krippendorff (2016) distinguishes between *replicability* – that is, the property that repeated measurements yield consistent results – and *reliability* – that is, the property that replicable and valid measurement can be relied upon to yield meaningful insights. In our paper, we stick to the more common usage wherein reliability includes both of said aspects.

[2] Importantly, random error does not presume that coders choose categories at random when they are undecided about a classification decision (Krippendorff, 2016), but merely that there is a component of disagreement that can be attributed to random processes, such as confusing cells in a spreadsheet, or hitting the wrong key.

of disagreement in textual measurement, we introduce a notion of *valid disagreement*, a form of replicable disagreement that validly captures relevant information about the textual content, and must be distinguished from replication failures that threaten reliability. Specifically, we distinguish between three key forms of meaning multiplicity that result in valid disagreement: In the first variant, meaning multiplicity arises from a scarcity of cues in the text about its intended meaning, raising *ambiguity*. In the second variant, meaning multiplicity arises from an abundance of textual cues pointing at multiple, co-present meanings, raising *polysemy*. In the third variant, meaning multiplicity arises not from the textual meaning itself, but from the presence of *interchangeable* opportunities for recording this meaning in the coded categories. We demonstrate that valid disagreement is widespread in many key domains of textual analysis in the social sciences, and defies adequate treatment within the confines of existing content analysis protocols. Discussing the problematic implications of failing to distinguish valid disagreement from measurement error, finally, we identify key properties of valid disagreement that enable us to recognize and address its impact on textual measurement.

2. Reliability in content analysis

Content analysis arguably constitutes the primary methodological contribution made by communication studies to the toolbox of social scientific research (Riffe et al., 2019). Since its inception, methodologists have recognized the need to ascertain the intersubjective quality of textual measurement. Key to this challenge is the fact that the property measured in textual material – its information or meaning – is not strictly speaking included in the text, but arises from the text by virtue of it being read and interpreted by a human reader (Krippendorff, 2004). Manifest signals within the text merely evoke conceptual categories presumably known to the reader and suggest specific relations among them (van Gorp, 2010), which enable the reader to infer an author's communicative intentions, and often copious additional information (Barthes, 1970; Franzosi, 1990). Consequently, it is always possible that the same text is understood in different ways by different readers, or by the same reader at different times (Eco, 1979; Krippendorff, 2004).

Content analysis, as a method, is paradigmatically oriented toward the intersubjective measurement of textual meaning, relying on categories that are imposed upon the text by the researcher (Krippendorff, 2008). Coders need to determine the presence of deductively defined meanings based on explicitly laid-down coding instructions and the text. Classification thus needs to remain independent of the intentions of the author – a principle that enables the measurement also of contents that were not intended or considered (e.g., abstract classifications such as frames or narrative schemata), or might even be denied by the author (e.g., relying on telling omissions, subtly encoded worldviews, or offensive implicatures). If coding instructions occasionally require coders to appraise textual evidence of authors' intentions, they do so not in order to reconstruct intended meanings, but to decide on a given classification defined in the codebook. Likewise, and in contrast to reception studies, content analysis must not depend on what readings are eventually actualized by audiences. Such meanings are inevitably co-shaped by diverse contexts, readers’ motivations and salient beliefs, and thus lack the required intersubjective quality (Fiske, 1979; Hall, 1980). Whenever researchers instruct coders to assume specific reading perspectives, they do not aim to inductively find diverse meaning potentials, but determine whether a deductively defined meaning is present. Understood in this way, meaning multiplicity arises not from common disagreements among authors and audiences over textual meanings or meaning potentials, but is better understood as one antecedent of such disagreements: It describes an intersubjective quality of the text itself, which exists whenever its meaning is not sufficiently constrained to enable a confident, unique decision on the applicability of one or multiple classifications.

Of course, most texts studied in social scientific textual research are sufficiently communicative to ensure that some degree of consensus can be achieved on their meanings (Wilson & Sperber, 2012), even if some inevitable ambiguities remain. However, in content analysis, coders frequently need to recognize meanings that were not necessarily intended or even considered by the authors of classified texts (especially, at the level of conceptual abstraction required by the content analyst). In fact, the authors of analyzed texts might even aim to disguise specific meanings or enable diverse readings (e.g., Baden & Sharon, 2020; Bavelas et al., 1990). In such cases, texts are unlikely to offer much guidance that facilitates the coders’ task (Oleinik et al, 2014).

To enable replicable classification despite the innate openness of texts and their frequent under-specification of measured meanings, content analysts need to offer considerable amounts of explicit guidance for the coding process. Codebooks contain detailed definitions, aiming to harmonize coders’ conceptual understanding of the coded constructs and their conceptual boundaries; they define interpretation rules and establish specific reading perspectives; they provide contextual cues and knowledge; and they lay down concrete inferencing rules and indicator sets to ensure that coders arrive at consistent categorizations of the text. Having thus constrained any idiosyncrasies in coders’ readings and classifications of the text, independent coders’ capacity to replicate the same classification decisions serves as the key test of the capacity of a codebook to enable reliable measurement.

Recognizing the critical importance of intersubjective measurement in content analysis, intense debates have unfolded over which coefficient is most suitable for a statistical evaluation of inter-coder reliability, and what levels of replicability are required to consider measurement sufficiently reliable to support scientific inferences. With regard to the choice of coefficient, most controversy has focused on how to best account for “chance agreement”, that is, the probability that coders agree on a classification in the absence of harmonizing coding rules (Krippendorff, 2011; Feng, 2012).[3] While several normalizations have been proposed, there is

[3] The key controversy concerns by which proportions coders would be expected to select which classification in the absence of a common understanding of the coding task. Accordingly, the interpretation of chance agreement (typically expressed by a replicability coefficient of zero) varies somewhat between coefficients. Chance-corrected replicability scores can be generally understood as a measure of alignment between different coders’ interpretations of the coding task (zero means that their choices are uncorrelated, possibly beyond a shared awareness of marginal class frequencies; negative values indicate a systematic dis-alignment).

growing consensus – at least in communication research – on Krippendorff's α as the measure of choice (Lovejoy et al., 2016), owing to its desirable behaviors (for a review see Krippendorff, 2008; Hayes & Krippendorff, 2007). With regard to the level of required agreement, different rules of thumb exist (Lovejoy et al., 2016), ranging from Landis and Koch's (1977) fairly lenient recommendations (permitting scores as low as $\kappa > 0.4$; see also Fleiss et al., 2003) to Krippendorff's (2008) rather strict cutoff value of $\alpha > 0.8$,[4] with 0.67 offering a widespread compromise (Craggs & Wood, 2005). Using simulation data, Geiss (2021) recently reopened this debate, demonstrating that for larger sample sizes, comparatively weak reliability measures may still suffice under certain circumstances, while high levels of agreement are required for detecting subtle patterns in small-scale data. Also in artificial intelligence research, recent advances in "weak supervision" point to the possibility of extracting an informative signal even from low-quality (i.e., not very reliable) annotations, provided that the remaining disagreement can be assumed to be relatively unstructured (Reidsma & Carletta, 2008; Ratner et al., 2018).[5] Despite the considerable methodological attention dedicated to the evaluation of reliability, however, to date no contribution has questioned the practice of equating reliability with inter-coder agreement, or its underlying assumption that classification must be definite and unique.

At present, all available measures of inter-coder reliability build on the assumption that any observed disagreement in classification indicates measurement error, while valid measurement must necessarily converge onto exactly one classification decision.[6] Of course, most content analysts have encountered classification decisions that were exceedingly hard to determine (e.g., Weber et al., 2018); however, the universal response to such challenges is to further narrow the coding instructions, pushing coders to agree. In the following, we will argue that forcing coders toward agreement at all costs runs the risk of misrepresenting the multiplicity of meanings available in the text, buying reliability at the cost of validity.

If meaning multiplicity is an inherent quality of textual communication, it follows that some classification decisions may not be mutually exclusive or may be impossible to make with definite confidence. Specifically, there are at least three forms of meaning multiplicity (see also Ceccarelli, 1998, for a related distinction) that challenge inter-coder reliability in conventional content analysis. First, texts frequently underspecify meaning; they provide scarce information about measured categories, for example, when categories target meanings that were of little or no concern to the author's communicative intention, or were deliberately kept am-

[4] However, Krippendorff (2004; 2008) expressly acknowledges that different standards may be applicable under different circumstances. Many scholars have warned that there are "no magic threshold[s]" (Craggs & Wood, 2005, p. 294), urging scholars to evaluate replicability in the context of a given coding task and application (Geiss, 2021).

[5] This may, for instance, be the case in crowd coding, where many coders' individual biases each contribute only little to measurement error. By contrast, computer-classified data are generally unsuitable for such uses, since misclassification by computational tools is never unstructured.

[6] This is ironically even the case for those heterogeneous residual categories often used to treat instances that cannot be confidently classified – instances for which "other" must then be chosen as definite and unique, "correct" classification (see Krippendorff, 2011, for a discussion of challenges caused by default categories).

biguous (Bavelas et al., 1990; Eisenberg, 2014). Second, texts sometimes invite or even require readers to construct multiple meanings. This is especially common in culturally rich texts whose communicative meaning arises from the collision or convergence of different modalities, intertexts, and connotations (Boxman-Shabtai, 2020; Fiske, 1989). Third, even if the communicative intention of a text is fairly clear, there may still be multiple, interchangeable ways for recording this meaning into coded categories. In the following, we will discuss each variant of meaning multiplicity in turn, examining how each cause irreducible, valid disagreement.

3. Ambiguity: Meaning multiplicity due to insufficient information

The first type of meaning multiplicity arises from basic communication norms: Following H. Paul Grice's (1975) maxim of quantity, communicators typically aim to make their contributions "as informative as is required (for the current purposes of the exchange)," but not "more informative than is required" (p. 44). Any information that is unnecessary to achieve an author's communicative goals can be omitted. Especially in real-time and interactive communication, rich information is contextually available, so restating it would be redundant.[7] Frequently, key information required for classification in content analysis is absent simply because it is not required to convey the author's communicative intention. Besides efficiency-driven omissions, authors may also deliberately underspecify the meaning of their contributions. For instance, authors may employ strategic ambiguity to enable different audiences to read diverse preferred meanings into the text (a common practice in political discourse; Eisenberg, 1984; Friedman & Kampf, 2014), or to avoid sanctioning while expressing controversial or risky meanings (Baden & Sharon, 2020; Bavelas, 1990). Thus, ambiguity stemming from information scarcity can severely constrain the range of categories that can be confidently determined in content analysis.

Demonstrating the characteristic challenges that arise from textual ambiguity, in one project concerned with how future events are discussed in public discourse (Tenenboim-Weinblatt et al., 2022a; 2022b), we encountered the following tweet by U.S. data journalist Nate Silver, referring to the 2016 presidential election:

> *There's a certain type of Democrat I talk to who seems determined not to get their hopes up that Trump will lose, which is an understandable reaction. But sometimes that morphs into a belief that it's savvy to think Trump will win/naive to think he *could* lose, when it isn't.*

While it is not part of the author's primary communicative intention to specify who will win – one of our coded categories – the tweet does bear upon this question, and Silver clearly has an opinion about it (marked by his evaluation: "it isn't"). Yet,

[7] This is also true to some extent for news discourse, which structurally builds upon and continues preceding news, such that individual news items rarely explicate all the information that is required to understand their meaning (Baden, 2018). That said, such omissions are less common in texts that are expected to be read by unspecified audiences over an extended range of time, as this mode of consumption limits authors' capacity to anticipate readers' available knowledge, and requires the text to explicate key points in order to maintain control over intended meanings.

what does he predict, and how should this prediction be coded? Does Silver anticipate a likely Trump defeat, or does he merely assert that all options are still on the table? Similarly, the implied prediction remains ambiguous in the following tweet by U.S. news columnist Thomas Friedman, published during the 2020 presidential election campaign: “Trump’s going to get re-elected, isn’t he?”. While the first part of the tweet suggests a definite outcome, the second part seems to disagree. Following Gricean communication norms, the appended question would be excessive if the answer was already given, suggesting that Friedman might exactly *not* predict a Trump re-election – which, owing to the binary U.S. party system, might amount to predicting a Biden win, or at least some non-trivial probability of it. For both tweets, coding a predicted possible, or even likely Trump defeat constitutes one plausible reading, but in neither case is this classification definite.

Ambiguity increases further when social sanctioning motivates authors to avoid fully specifying their intended meaning. In another project, which aimed to recognize references to conspiracy theories in news users’ commentary (Baden & Sharon, 2020), many posts relied on contextual knowledge to allude to possible conspiracy theories: For instance, underneath an article covering the early stages of the Covid-19 pandemic on the Israeli news site Walla!, one user mused: “Isn’t it strange that the doctors and nurses haven’t yet been infected and died?”. While we can recognize the presupposed claim (that they did not get infected) as contradicting conventional knowledge, one common criterion for classification as conspiracy theory (Birchall, 2006), this is not evident from the text alone. In this case, another user helped by explicating the epistemic conflict, responding “In China, they died,” prompting the first user to double down: “Doctors and nurses and everyone who works with clients has been immunized.” Yet, ambiguity remains: While the “strange” in the first comment marks a call for (alternative) explanations and the use of “everyone” may signal the presence of some hidden, powerful agent working toward some sinister end (the “conspiracy”), this is only one available reading of the exchange. Only if coders assume that the user intended to reference a specific heterodox account of the beginning pandemic (which viewed the virus as a bioweapon deliberately released by China to weaken the West) can the comment be classified as a conspiracy theory reference.

Ambiguity is widespread in communication, and has long been discussed by both theorists and methodologists. While there is long-standing consensus in qualitative textual analysis that competing meaning potentials need to be explicitly addressed, explored and validated (Carey, 2008; Morley, 1980; Liebes & Katz, 1990), *quantitative* textual analysis has mostly tried to evade or subdue ambiguity (Krippendorff & Craggs, 2016). On the one hand, content analysts have responded by retreating to focus on relatively manifest claims (Bolognesi et al., 2016), which can be coded with higher confidence, discarding more latent meanings. Unfortunately, as we have demonstrated, manifest contents may often present an inaccurate measurement of textual meaning. Moreover, the strategy inevitably neglects large parts of the information whose relevance to coded categories is evident, even if its specific meaning is undetermined. This is especially costly in the study of controversial and sanctioned meanings, as it systematically undermeasures any contents that attempt to avoid or bridge controversy or mitigate the author’s commitment. In the mentioned study of conspiracy theory references, only a miniscule fraction of posts qualified unambiguously, while the vast majority were ambiguous (Baden & Sharon, 2020). On the other hand, content analysts have attempted to subdue ambiguity by adding assumptions about readers' available knowledge and reading perspective to the coding instructions (Oleinik et al., 2014). For instance, we might assume a paranoid reader who perceives conspiracies wherever possible, enabling us to code ambiguous cases as long as this is one viable interpretation;[8] or we could assume a mainstream reader, who will only recognize fairly blunt conspiracy theory references (although it remains almost impossible to define just how blunt is blunt enough, unless one retreats again to a reliance on manifest indicators). While this solution may be productive for applications that are specifically interested in the textual meanings presented to specific kinds of readers (e.g., Baden, 2018; Liebes & Katz, 1990), it also adds systematic bias to the measurement, which needs documentation and justification. Moreover, it requires quite bold, and typically untested assumptions about how coded texts would be read, and systematically overstates the degree of explicitness of measured meanings.

4. Polysemy: Meaning multiplicity due to excessive information

The second type of meaning multiplicity arises not from the scarcity, but from the abundance of information encoded in the text and its implication for various forms of decoding (Hall, 1980). Polysemy is typically additive, aggregating meaning from the simultaneous co-presence of codes, intertextual references, and modalities that influence and infuse one another (Boxman-Shabtai, 2020; Eco, 1997). The prevalence of polysemy varies by genre: technical and administrative texts tend to avoid polysemy, while many forms of cultural communication are saturated with it (Fiske, 1987). For some genres, polysemy is almost constitutive: Notably, several forms of humor are funny exactly because of a clash between incongruent scripts, an appreciation of which requires recognition of distinct, co-present meanings (Boxman-Shabtai & Shifman, 2014; Raskin, 1984). Some forms of polysemy rely on an interplay between literal meaning and connotation (e.g., "thoughts and prayers") and play with lexical similarity (e.g., Israeli protesters' labeling of PM Benjamin Netanyahu as "Crime Minister"). Others rely on metaphors and analogies (e.g., South African comedian Trevor Noah comparing U.S. president Trump to a cliché African dictator; anti-immigrant politicians referring to refugees as a tidal wave), or make use of intertextuality as a means for importing additional meanings (e.g., Chan et al.'s [2020] use of "the Babel problem" in multilingual AI imports additional, intriguingly relevant Biblical meanings about the hubris of playing god).

In some cases, the primary communicated meaning arises directly from the collision of multiple meanings, as in the case of irony, which contrasts literal against pragmatic meaning (Gal, 2019), or of "stereotypical overload" in ethnic humor, which criticizes stereotypical perceptions by their demonstrative exaggeration (Boxman-Shabtai & Shifman, 2014). In other cases, polysemy plays an auxiliary

[8] Unfortunately, conspiracy theories are available as possible readings for an astounding wealth of texts.

role, adding nuance or suggesting specific perspectives for interpretation, sometimes providing clues that undermine the main message ("semantic slip"; e.g., Fiske, 1987). What all variants have in common is that the text expressly invites multiple, co-present readings that jointly contribute to the textual meaning in ways that cannot be captured by either reading alone. Consequently, in the case of polysemy, valid disagreement focuses less on whether a given category is present in the text, and more on which category, or which categories best capture its meaning. Is the famous campaign slogan "It's the economy, stupid" about economics, about politics, or both – or isn't the point exactly that both topics can't be separated?

Examples abound in an ongoing study investigating expressions of political critique related to the Israeli-Palestinian conflict on TikTok, a platform expressly designed for the memetic re-use and intertextual re-contextualization of pop-culturally relevant meanings (Literat et al., 2022; Zulli & Zulli, 2022). In one instance, Palestinian protesters were shown raising a sign inscribed "We cannot breathe since 1948," drawing an analogy between the final words uttered by George Floyd, a member of the African American community murdered in 2020 by Minnesota police, and the lives of Palestinians in the presence of the state of Israel. Numerous interpretations are supported by this reference, most of which reinforce one another, while remaining conceptually separate. The sign evokes a parallel between Palestinian national activism and the Black Lives Matter movement protesting U.S. police violence against black community members; Israel (or Israeli security forces?) is likened to racist U.S. police officers, and Palestinians are likened to either the (murdered) George Floyd or the victimized black community as a whole. While it is easy to classify the video as critical of Israel (the intended, convergent meaning shared between the available readings), should the video also be coded as critical of the U.S., or oppression on a global scale? Does this form of criticism deny Israel's right to exist (as might be argued based on the reference to 1948, Israel's year of independence, as opposed to 1967, the year more closely associated with its occupation of the Palestinian territories)? In each case, possible grounds for multiple classification decisions are in plain evidence.

In another video, set against the soundtrack of a dialogue taken from the "Mean Girls" movie, various meanings are co-present in a rather disorienting manner. The audible soundtrack revolves around an argument between Cady Heron (Lindsay Lohan), the movie's lead protagonist, and her friend Janis Ian (Lizzy Caplan) who accuses Heron of betraying her real friends in favor of villain character Regina George and her clique ("the plastics"). Against this soundtrack, a young man plays both roles, superimposing the conflict script in Hebrew subtitles and textual labels that cast Israeli-born actor Natalie Portman as Cady heron, and the state of Israel as the betrayed friend. Ridiculing Portman's critical position toward Israel (as phony, both through the analogy to the altercation in Mean Girls and the visual performance), and explicitly criticizing her as "dirty little liar" (in both text and audio), the video simultaneously addresses several interlinked discursive arenas. On one level, it accuses Portman of dishonesty and lacking patriotism; on another level, it casts U.S. Jews as out of touch and overly concerned with appearances; and on yet another level, it likens the entire controversy to some form of soap opera defined by superficial egoisms and personal grievances. Once again, classification is easy where the text's various meanings converge (here, expressing a pro-Israel position), but otherwise polysemic: For instance, is the demand for patriotic allegiance sincere or does the embedding in the Mean Girls context trivialize or even undo the presented criticism? Does Portman represent only herself, or does she stand representative for the community of Israel-critical U.S. Jews? Multiple competing classification choices validly capture different facets of the complex meaning.

While some scholars have highlighted the difficulties emerging from systematically coding polysemic constructs in the context of humor (Boxman-Shabtai & Shifman, 2014; Nissenbaum & Shifman, 2020), there is still very little work that focuses on the implications of polysemy for quantitative social science research. In a rare pertinent contribution, Krippendorff and Craggs (2016) acknowledge that "there are many situations in which units are more naturally described in terms of multiple values" (p. 182). Rejecting the common strategy of "impos[ing] coding restrictions on multi-valued phenomena, instructing coders to record only the most prominent of several applicable attributes," the authors propose a solution wherein multiple valid codes are recorded simultaneously (checking all classes that apply). Still, they note, "all contiguities intrinsic to multi-valued accounts of phenomena are lost" (p. 186), including any meanings that arise only from the collision of co-present meanings. Especially in heavily intertextual and multimodal genres of text, such classification runs the risk that some categories are almost always present (for instance, in political discourse, relatively few texts contain no positive or negative evaluations at all), rendering multiple classification relatively useless. In addition, the strategy multiplies the number of required coding decisions, a questionable strategy for improving reliability especially considering that "the act of recognizing whether a variable applies tends to be far more unreliable than distinguishing among the values of an applicable variable" (Krippendorff & Craggs, 2016, p. 186). That said, the alternative conventional solution of adding heterogeneous "other" categories (e.g., "ambivalent") to capture complex meanings effectively gives up on polysemy by excluding it from subsequent analyses (Krippendorff, 2011).

5. Interchangeability: Meaning multiplicity due to modular coding

A final variant of meaning multiplicity arises not from the text alone, but from the attempt to map textual meanings onto those coded categories offered by the coding instructions. While both ambiguity and polysemy are liable to raise such difficulties, uncertainties in the coding of textual meanings may remain even if the textual meaning is fairly clear and relevant to the intended classification. To see how, it is useful to recognize that the measurement of complex textual meanings frequently requires breaking down abstract conceptual categories into their constituent components (Franzosi, 1990). For example, in the project aiming to measure future-oriented scenarios in public discourse (Tenenboim-Weinblatt et al., 2022a; 2022b), attempting measurement at the level of entire predicted scenarios is essentially infeasible. As predictions may concern virtually any conceivable subject matter and are further differentiated by their estimated probability, desirability, and other factors, any attempt at capturing all relevant variations inevitably generates a huge taxonomy of complex, overlapping categories that are impossible to code in an intersubjectively replicable manner.

Instead, a common strategy is to break complex, multidimensional constructs down into their constituent parts, which can be coded separately, with much superior reliability (e.g., Matthes & Kohring, 2008). Many constructs commonly measured in social scientific textual research support such an approach: For instance, evaluations can be classified based on their evaluative tendency (positive/negative) and evaluative standard (e.g., morality, aesthetics, functionality; Baden & Springer, 2014; Weber et al., 2018); attributions can be measured by separately recording the object and the attributed quality (e.g., Fridkin & Kenney, 2011); frames are often measured by identifying their interdependent frame elements separately (e.g., Entman, 1993; Matthes & Kohring, 2008); and the list continues. The catch, of course, is that in such modular coding, classification decisions become interdependent. Such issues typically arise whenever measurement depends on utterances that address some relationship between different constituent parts, especially when this relationship is dissociative or negated (e.g., "he isn't all that smart" implies that he is dumb; "she got fired" implies that someone else fired her, and that she is now unemployed; Franzosi, 1990).

We found election-related forecasts to frequently raise interchangeable classification options. For example, U.S. politician Stacey Abrams tweeted in the course of the 2020 presidential elections:

> *Voter fraud is a myth that is perpetuated by Donald Trump to hide the fact that he knows that if there is full participation, he will likely not win.*

To capture the presented projections, our coding scheme measured a) who or what a projection was about; b) if it was about a contender, whether it predicted a win (including successes on the way) or a defeat (including setbacks on the way); and c) what probability was attributed to the presented outcome. While it is clear what Abrams predicts here – that Trump "will likely not win" – there are two interchangeable ways of validly recording this information: If we attribute the negation to the win, we need to code a: Trump; b: defeat ("not win"); and c: likely. However, if we attribute the negation to the probability statement, we obtain a: Trump; b: win; and c: unlikely ("likely not"). Both variants are valid and accurately capture the projection.

Things get murkier yet when statements mention both contenders, such as U.S. senator Bernie Sanders' comment that "I do not believe that we will defeat Donald Trump with a candidate like Joe Biden." Here, we get four plausible options: We can code (1) an unlikely ("I do not believe") Biden win ("we will") or (2) Trump defeat ("defeat Trump"); or (3) a likely Trump win, or (4) Biden defeat.

In practice, coding interchangeability is typically solved by adding formal coding rules to the instructions (Krippendorff, 2004). In our case, for instance, we ruled that if a prediction involved both one candidate's win and another's (thereby logically implied) defeat, only the winning prediction should be coded; we decided to follow the text in deciding whether one outcome was marked as likely or its inverse as unlikely; and we added a rule to the effect that, if a negation could be read equally well as part of the predicted state ("not win") and as part of the probability ("likely not"), it should be attributed to the probability. In our latter example, these rules lead us to prefer variant 1 (unlikely Biden win), enabling confident and reliable classification. As a side effect, however, even slight, semantically irrelevant variations in the phrasing of substantially identical predictions thus require different classification choices, introducing meaningless variance into the data (Franzosi, 1990). Moreover, especially in the classification of complex constructs, such arbitrary rules quickly multiply in ways that create new uncertainties (e.g., which rules take precedence over others), generating many new opportunities for human classification error.

Finally, all three types of meaning multiplicity – ambiguity, polysemy, interchangeability, summarized in Table 1 – occasionally coalesce, creating a truly complex challenge. As an example, when U.S. economist and New York Times columnist Paul Krugman predicted that "it will be almost impossible for Trump to win reelection legitimately," the seemingly minor addition of "legitimately" creates multiple problems at once. In terms of ambiguity, Krugman's vagueness about what kind of illegitimate means could still enable a Trump win elude our outcome category, which is premised on the assumption of a legal election. Specifically, this complicated our probability classification: If we assume the expression to refer to objectionable but legal means (e.g., intimidating voters; disinformation campaigns), Krugman still predicts that Trump can win and be legally elected (we would code this as "remotely possible"); but if we assume illegal means, such as fraud or violent insurrection, Trump wouldn't actually win, and we would code a Trump win as "impossible/very unlikely"). Depending on the "preferred reading" (Hall, 1980), which is shaped by differences in ideology and public discourse (e.g., left-leaning U.S. media cast Trump as quite capable of fraud), different classification decisions follow.

At the same time, "legitimately" also creates polysemy, as it raises two competing, but co-present scenarios: One wherein Trump loses; and one wherein he seeks to win by illegitimate means. While the former scenario comes with an explicit probability estimate ("almost impossible"), the latter does not – so if we consider the scenario of an illegitimate Trump victory, we would need to treat the outcome as "possible," which was our residual category for scenarios with unqualified probability. Importantly, both meanings are clearly present, in the sense that either viable classification choices (Trump/win/very unlikely; Trump/win/possible) misses a valid part of the information.

On top of both the ambiguity and polysemy, finally, we also get interchangeability, since "almost impossible ... to win" could count either as very likely defeat, or as very unlikely victory. In the present case, our reliance on the text would point us toward the latter option; however, had Krugman said "avoid defeat" in place of "win", neither our reliance on the text, nor our preference rule for qualifying probabilities before outcomes could resolve the dilemma. For this segment alone, therefore, we could obtain at least five valid alternative classifications in just two variables:

- if we assume legal means (ambiguity), *remotely possible / victory* or (interchangeability) *very possible / defeat*

- if we assume illegal means (ambiguity), *very unlikely / victory* or (interchangeability) *very likely / defeat*
- if we additionally consider the possibility of winning illegitimately (polysemy), *possible / victory*

Table 1. Three types of meaning multiplicity that cause valid disagreement

	Ambiguity	Polysemy		Interchangeability
Meaning multiplicity	text does not sufficiently specify coded meaning	text cues multiple co-present meanings		meaning maps in multiple ways onto coded categories
Valid disagreement	specific category coded as present or absent	varying selections of multiple applicable categories		alternative selections of interdependent categories
Conventional treatment	focus on manifest, well-specified contents only	fix interpretation perspective and inference rules	permit multiple classification	add arbitrary, formal classification rules
Resulting measurement biases	misses ambiguous, hedged instances	misses alternative or diverse readings	loses meanings that arise from interaction	introduces artifactual variance
Impact on measurement error	arbitrary explicitness threshold increases chances of error	multiplication of decisions increases chances of error		multiplication of rules increases chances of error

6. Valid disagreement in content analysis

As we have demonstrated, the presence of meaning multiplicity is not a matter of interpretation, or a result of insufficiently specific coding instructions, but a valid property of the text. It arises from common practices in language use, which undermine efforts at deductive text classification. While each variant of meaning multiplicity arises from distinct causes, all three of them invite disagreement among coders that validly reflects the multiplicity of meanings supported by the coded material. To date, such valid disagreement mostly shows up as disagreement in reliability measures, where it is treated as measurement error and motivates content analysts to further increase agreement among coders (Krippendorff, 2004). Of course, it is generally possible to increase agreement, even if such gains come at a cost: An overreliance on manifest contents often misrepresents textual meaning and systematically misses specific kinds of communicated meanings, especially those whose expression is socially sanctioned; fixing interpretation rules and constraining coders' perspective effectively shifts the measurement level from recording the contents of texts toward texts' presumed reception among specific audiences; multi-valued classification erases the meaning that arises from the co-presence of distinct categories, while treating polysemic meanings as "other" or "ambiguous" effectively removes them from the analysis – again focusing the analysis on a very specific subset of communicated meanings; tie-breaking rules are liable to introduce biases (e.g., overmeasurement of preferred-coded variants) and artifactual variance (e.g., coding the substantively same meaning differently depending on the formulation); and all of the above expand the set of interdependent classification rules, adding complexity and multiplying opportunities for human error. Of course, some of these interventions may be legitimate and even productive under certain circumstances, provided that they are rendered transparent, justified, and their implications for obtained measurements are discussed. However, while agreement can in most cases be forced, we hope that we could convince the reader that trying to reduce meaning multiplicity to definite, unique classification decisions potentially erodes the very validity of measurement that the pursuit of reliability was intended to enable.

In addition, the practice of conflating valid disagreement with measurement error has an even more dramatic impact upon social scientific textual research as a field. Confronted with measurement problems liable to involve meaning multiplicity, researchers frequently find themselves torn between three distinctly undesirable avenues. First, they can attempt to measure affected concepts, confronting the likely possibility of failing to reach conventional reliability standards. As valid disagreement depresses all established reliability metrics, researchers face likely rejection of their research reports, and may be forced to fall back on inferior publication venues.[9] Second, researchers may try to subdue valid disagreement by resorting to any combination of the agreement-raising strategies discussed above, potentially at the expense of measurement validity. Third, researchers may eschew problematic concepts altogether, retreating to measuring contents that are less affected by valid disagreement – notably, mentions (e.g., of named entities, constructs, which can usually be located with a high degree of intersubjectivity) and highly abstracted variables (e.g., topicality, sentiment, stance, which typically leave many redundant traces in text and are thus recognized more easily). In our own work, we have faced similar challenges at repeated occasions, struggling to strike a defensible compromise that achieves sufficient agreement without sacrificing too much valid complexity.

Regardless of the avenue chosen, however, the primary loss is that methodologically ambitious, valid textual research involving key concepts in the social sciences becomes exceedingly difficult to place in the central venues of the field. Instead, this space threatens to be occupied by textual research that relies on crude or only vaguely related, but replicable measures, passed off as proxies for more demanding constructs (Bolognesi et al., 2016) – be that sentiment measures standing in for evaluative opinions, thematic classifications passed off as frames, or other creative operationalizations.[10] In the final part of this paper, we will therefore

[9] Another, potentially unethical response might be to try and get away with reporting inferior (chance-uncorrected) replicability coefficients or none at all, both still common practices in the field (Lovejoy et al., 2016).

[10] This tendency is especially pronounced in computational text analysis (Baden et al., 2022)

propose a different avenue, based on the premise that disagreement raised by meaning multiplicity contains valid information that should not be discarded.

7. Revising reliability to account for valid disagreement

If meaning multiplicity is real (Barthes, 1970: Eco, 1979) and affects content analysis in ways that can't be subdued by conventional strategies, we should meet it head-on, measuring and modeling its impact. One immediate, valuable approach is to qualitatively analyze coding disagreements, gauging the extent to which divergent classification decisions represent alternative or additional, valid readings of the material, or are better understood as coding error (see, for example, Boxman-Shabtai & Shifman, 2014). Building upon extant research in discourse studies (Kintsch & van Dijk, 1983; Wilson & Sperber, 2012), Baden (2018) has proposed a conceptual framework for modeling the way in which contextual knowledge contributes to resolving textual ambiguity and polysemy.

Quantitative strategies are also readily available. Key to this shift in perspectives is the recognition that valid disagreement is replicable: Although textual ambiguity creates disagreement, especially when coders are forced to choose between multiple valid readings, such disagreement follows a predictable structure. This is most obviously so for interchangeability, which arises from the availability of competing options for recording the same meaning. For instance, all our cases of mapping uncertainty in future scenarios essentially oscillate between two specific, paired solutions that encode equivalent meaning (i.e., a likely defeat is the same as an unlikely victory). Regardless of the type of meaning multiplicity, valid disagreement is text-specific: The same texts consistently raise the same disagreements, regardless of the identity of coders. Accordingly, valid disagreement is statistically similar to coder bias, which results in disagreements that are consistent for the same coder, regardless of the classified text (Krippendorff, 2008; Reidsma & Carletta, 2008). Just as coder bias can be measured by estimating the extent to which the identity of the coder predicts differences in the coding, valid disagreement can be measured by estimating the extent to which specific texts, kinds of texts, or even textual indicators predict variation in the coding. If to measure coder bias, we have the same coders classify several texts (Krippendorff, 2004; Riffe et al., 2019), having the same texts classified by a number of coders (e.g., via crowd coding; Krippendorff, 2021) should reveal the extent to which textual properties are responsible for coders' failure to agree, indicating valid disagreement. Adjusting conventional replicability measures by discounting the prevalence of valid disagreement from the observed disagreement may thus offer a viable strategy for redeeming measurement efforts affected by meaning multiplicity.

Especially for research that taps into meanings heavily affected by ambiguity and/or polysemy, it may be wise to have many, if not all texts classified by multiple coders. If two coders per text already offer valuable insights into the relative prevalence of unique and polysemic texts (provided that a smaller share of texts is coded many times to gauge the ratio of valid versus erroneous disagreement), more coders yield incrementally more information. For instance, observing the same text being coded five times not only offers valuable insights into the "decidability"[11] of the coding task, but also into the relative prominence of multiple or alternative meanings (Baden, 2018). Still more can be learned from having the material coded by diverse populations of coders (e.g., representing different identity groups), revealing how differences in cultural context affect the textual measurement (e.g., Hallinan et al., 2021).

Finally, there may be a valuable cross-fertilization between manual and computational text classification. While human coders naturally struggle with meaning multiplicity, machine classifications are weighted and non-definite by design: Regardless of the selected algorithm, any classification tool assigns continuous weights and probabilities to each category, which are only converted into binary classification decisions in the end. While machine classification remains far from being capable of resolving ambiguous textual meanings, and is reliably fooled by polysemy, it may be useful for detecting which texts seem most likely to permit multiple or competing classifications (and potentially, what textual structures are likely responsible; Boxman-Shabtai & Shifman, 2014). Providing human coders with machine-rated weights may thus help sensitize them to the presence of potential meaning multiplicities.

In conclusion, we have argued in this intervention that textual ambiguity, polysemy and interchangeability are real and pervasive, and result in systematic, replicable disagreement in content analytic measurement. This valid disagreement records meaning multiplicity, a relevant, often deliberate property of the text (Boxman-Shabtai, 2020), and must not be confounded with measurement error. None of the three common variants of valid disagreement identified in this paper can be adequately resolved by forcing classification toward definite, unique solutions; doing so, we argue, achieves gains toward a misunderstood notion of reliability, at the expense of measurement validity. At present, ignoring valid disagreement in social scientific text analysis unduly depresses conventional reliability metrics, threatening the viability of textual measurement for critical constructs and research sites affected by meaning multiplicity. Instead, valid disagreement should be regarded as a common property of textual measurement that content analysts need to consider.

To address this challenge, we have proposed a revision of existing understandings of intercoder reliability, considering the important implications of valid disagreement in textual measurement. A methodological debate is overdue about the contribution of textual meaning multiplicity to measurement challenges in content analysis. Numerous well-established insights from qualitative text analysis, of which we could only reference a few in this intervention, stand ready to be considered. As an immediate response, we need to separate valid disagreement from measurement error in our evaluation of reliability. By correcting intercoder reliability scores to count valid disagreement as reliable information, high-quality studies affected by meaning multiplicity should be able to cross conventional reliability thresholds,

[11] For this notion, we are indebted to Krippendorff's final work (2021), which introduces a measure of "decisiveness", i.e., the tendency of multiple coders to reach unanimity. However, as decisiveness is a function of the degree of meaning multiplicity present in the textual material, we prefer the notion of "decidability", which conceptualizes a property of the coding task applied to a certain kind of textual discourse, and not a property of the coders' measurement.

https://doi.org/10.5771/2192-4007-2023-4-305 https://www.inlibra.com/de/agb - Open Access - BY-NC-ND

or justify their failure to do so by documenting that part of the observed disagreement is valid. Conceptually, what we call for is a revision of conventional interpretations of reliability as agreement, and of disagreement as error. While erroneous disagreement remains a threat to reliable measurement, some disagreement in classification is based in real textual ambiguity; it is structured, it is replicable, and most importantly – it is valid.

Funding

This work was supported by the European Research Council, ERC StG 802990.

References

Baden, C. (2018). Reconstructing frames from intertextual news discourse: A semantic network approach to news framing analysis. In P. D'Angelo (Ed.), *Doing news framing analysis II: Empirical and theoretical perspectives* (pp. 3–26). Routledge.

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods & Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Baden, C., & Sharon, T. (2020). Blinded by the lies? Toward an integrated definition of conspiracy theories. *Communication Theory*, *31*(1), 82–106. https://doi.org/10.1093/ct/qtaa023

Baden, C., & Springer, N. (2014). Com(ple)menting the news on the financial crisis: The contribution of news users' commentary to the diversity of viewpoints in the public debate. *European Journal of Communication*, *29*(5), 529–548. https://doi.org/10.1177/0267323114538724

Barthes, R. (1970). *S/Z*. Editions du Seuil.

Bavelas, J. B., Black, B., Chovil, N., & Mullett, J. (1990). *Equivocal communication.* Sage.

Bolognesi, M., Pilgram, R., & van den Heerik, R. (2016). Reliability in content analysis: The case of semantic feature norms classification. *Behavioral Research*, *49*, 1984–2001. https://doi.org/10/3758/s13428-016-0838-6

Boxman-Shabtai, L., & Shifman, L. (2014). Evasive targets: Deciphering polysemy in mediated humor. *Journal of Communication*, *64*(5), 977–998. https://doi.org/10.1111/jcom.12116

Boxman-Shabtai, L. (2020). Meaning multiplicity across communication subfields: Bridging the gaps. *Journal of Communication*, *70*(3), 401–423. https://doi.org/10.1093/joc/jqaa008

Carey, J. W. (2008). A cultural approach to communication. In G. S. Adam (Ed.), *Communication as culture: Essays on media and society* (2nd ed., pp. 11–18). Routledge.

Ceccarelli, L. (1998). Polysemy: Multiple meanings in rhetorical criticism. *Quarterly Journal of Speech*, *84*(4), 395–415. https://doi.org/10.1080/00335639809384229

Ceccarelli, L. (2001). *Shaping science with rhetoric: The cases of Dobzhansky, Schrodinger, and Wilson.* University of Chicago Press.

Chan, C.-H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., . . . Althaus, S. L. (2020). Reproducible extraction of cross-lingual topics (rectr). *Communication*

Methods & Measures, *14*(4), 285–305. https://doi.org/10.1080/19312458.2020.1812555

Craggs, R., & Wood, M. M. (2005). Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, *31*(3), 289–296. https://doi.org/10.1162/089120105774321109

Eco, U. (1979). *The role of the reader: Explorations in the semiotics of texts*. Indiana UP.

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. *Communication Monographs*, *51*(3), 227–242. https://doi.org/10.1080/03637758409390197

Feng, G. C. (2012). Factors affecting intercoder reliability: A Monte Carlo experiment. *Quality & Quantity*, *47*, 2959–2982. https://doi.org/10.1007/s11135-012-9745-9

Fiske, J. (1989). *Reading the popular.* Routledge.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Wiley.

Franzosi, R. (1990). Strategies for the prevention, detection, and correction of measurement error in data collected from textual sources. *Sociological Methods & Research*, *18*(4), 442–472.

Fridkin, K. L., & Kenney, P. J. (2011). The role of candidate traits in campaigns. *The Journal of Politics*, *73*(1), 61–73. https://doi.org/10.1017/S0022381610000861

Friedman, E., & Kampf, Z. (2014). Politically speaking at home and abroad: A typology of message gap strategies. *Discourse & Society*, *25*(6), 706–24. https://doi.org/10.1177/0957926514536836

Gal, N. (2019). Ironic humor on social media as participatory boundary work. *New Media & Society, 21*(3), 729–749. https://doi.org/10.1177/1461444818805719

Geiß, S. (2021). Statistical power in content analysis designs: How effect size, sample size and coding accuracy jointly affect hypothesis testing – A Monte Carlo simulation approach. *Computational Communication Research*, *3*(1), 61–89. https://computational-communication.org/ccr/article/view/44

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: Vol. 3: Speech acts*. Academic Press.

Hall, S. (1980). Encoding/decoding. In S. Hall, D. Hobson, A. Love, & P. Willis (Eds.), *Culture, media, language: Working papers in cultural studies, 1972-79* (pp. 128–138). Hutchinson.

Halliday, M. (1978). *Language as social semiotic*. Arnold.

Hallinan, B., Kim, B., Scharlach, R., Trillò, T., Mizoroki, S., & Shifman, L. (2021). Mapping the transnational imaginary of social media genres. *New Media & Society*, *25*(3), 559–583. https://doi.org/10.1177/14614448211012

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods & Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Joseph, R. L. (2018). *Postracial resistance: Black women, media, and the uses of strategic ambiguity.* NYU Press.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Sage.

Krippendorff, K. (2008). Systematic and random disagreement and the reliability of nominal data. *Communication Methods & Measures*, 2(4), 323–338. https://doi.org/10.1080/19312450802467134

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods & Measures*, *5*(2), 91–112. https://doi.org/10.1080/19312458.2011.568376

Krippendorff, K. (2016). Misunderstanding reliability. *Methodology, 12*(4), 139–144. https://doi.org/10.1027/1614-2241/a00019

Krippendorff, K. (2021). A quadrilogy for (big) data reliabilities. *Communication Methods & Measures*, *15*(3), 165–189. https://doi.org/10.1080/19312458.2020.1861592

Krippendorff, K., & Craggs, R. (2016). The reliability of multi-valued coding of data. *Communication Methods & Measures*, *10*(4), 181–198. https://doi.org/10.1080/19312458.2016.1228863

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Liebes, T., & Katz, E. (1990). *The export of meaning: Cross-cultural readings of Dallas*. Oxford University Press.

Literat, I., Boxman-Shabtai, L., & Kligler-Vilenchik, N. (2022). Protesting the protest paradigm: TikTok as a space for media criticism. *The International Journal of Press/Politics*, *28*(2), 362–383. https://doi.org/10.1177/1940161222111748

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

Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, *58*, 258–279. https://doi.org/10.1111/j.1460-2466.2008.00384.x

Mikhaylov, S., Laver, M., & Benoit, K. R. (2011). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, *20*, 78–91. https://doi.org/10.1093/pan/mpr047

Morley, D. (1980). *The "Nationwide" audience: Structure and decoding*. BFI.

Oleinik, A., Popova, I., Kirdina, S., & Shatalova, T. (2014). On the choice of measures of reliability and validity in the content-analysis of texts. *Quality & Quantity*, *48*, 2703–2718. https://doi.org/10.1007/s11135-013-9919-0

Nissenbaum, A., & Shifman, L. (2020). Laughing alone, together: Local user-generated satirical responses to a global event. *Information, Communication & Society*, *25*(7), 924–941. https://doi.org/10.1080/1369118X.2020.1804979

Raskin, V. (1984). *Semantic mechanisms of humor.* Springer.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2018). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, *11*, 269.

Reidsma, D., & Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, *34*(3), 319–326. https://doi.org/10.1162/coli.2008.34.3.319

Riffe, D., Lacy, S., Watson, B. R., & Fico, F. (2019). *Analyzing media messages: Using quantitative content analysis in research* (4th ed.). Routledge.

Shifman, L. (2013). *Memes in digital culture.* MIT Press.

Tenenboim-Weinblatt, K., Baden, C., Aharoni, T., & Overbeck, M. (2022a). Affective forecasting in elections: A socio-communicative perspective. *Human Communication Research*, *48*(4), 553–566. https://doi.org/10.1093/hcr/hqac007

Tenenboim-Weinblatt, K., Baden, C., Aharoni, T., & Overbeck, M. (2022b). Persistent optimism under political uncertainty: The evolution of citizens' political projections in repeated elections. In M. Shamir & G. Rahat (Ed.), *The Elections in Israel, 2019–2021*. Routledge.

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., Gordon, A., Khooshabeh, P., Hahn, L., & Tamborini, R. (2018) Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, *12*(2-3), 119–139. https://doi.org/10.1080/19312458.2018.1447656

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. CUP.

Zhao, X., Feng, G. C., Ao, S. H., & Liu, P. L. (2022). Interrater reliability estimators tested against true interrater reliabilities. *BMC Medical Research Methodology*, *22*(232), 1–19. https://doi.org/10.1186/s12874-022-01707-5

Zulli, D., & Zulli, D. J. (2022). Extending the Internet meme: Conceptualizing technological mimesis and imitation publics on the TikTok platform. *New Media & Society*, *24*(8), 1872–1890. https://doi.org/10.1177/1461444820983603