Textual Indicators of Deliberative Dialogue

A systematic review of methods for studying the quality of online dialogues

Mr Alex Goddard[a], Professor Alex Gillespie[a,b]

[a]Department of Psychological and Behavioural Sciences, London School of Economics and Political Sciences, Houghton Street, London, WC2A 2AE, UK

[b]Department of Psychology, Oslo New University College, Oslo, Norway

Corresponding author:
Alex Goddard,
Department of Psychological and Behavioural Sciences,
London School of Economics and Political Sciences,
Houghton Street,
London,
WC2A 2AE,
UK
Email: a.j.goddard@lse.ac.uk

The Authors declare that there is no conflict of interest

Abstract:

High-quality online dialogues help sustain democracy. Deliberative theory, which predates the internet, provides the primary model for assessing the quality of online dialogues. It conceptualizes high-quality online dialogue as civil, rational, constructive, equal, interactive, and for the common good. More recently, advances in computation have driven an upsurge of empirical studies using automated methods for operationalizing online dialogue and measuring its quality. While related in their aims, deliberative theory and the wider empirical literature generally operate independently. To bridge the gap between the two literatures, we introduce *Textual Indicators of Deliberative Dialogue* (TIDDs). TIDDs are defined as text-based measures of online dialogue quality under a deliberative model (e.g., disagreement, incivility, justifications). In this study, we identified 123 TIDDs by systematically reviewing 67 empirical studies of online dialogue. We found them to have mid-low reliability, low criterion validity, and high construct validity for measuring two deliberative dimensions (civility and rationality). Our results highlight the limitations of deliberative theory for conceptualizing the variety of ways online dialogues can be operationalized. We report the most promising TIDDs for measuring the quality of online dialogue and suggest deliberative theory would benefit from altering its models in line with the broader empirical literature.

Textual Indicators of Deliberative Dialogue

A systematic review of methods for studying the quality of online dialogues

Humanity faces increasingly global problems requiring large-scale coordination. Online dialogues offer a public space where people can discuss issues of common concern. Deliberative theory argues that when online dialogues are high quality they maintain healthy democracies (Dahlberg, 2001; Friess & Eilders, 2015; Graham & Wright, 2014; Janssen & Kies, 2005; Sunstein, 2018). Advances in computation have led to an upsurge in empirical studies using quantitative text analysis methods for analyzing online dialogues (Lampe, 2013). Deliberative theory currently operates independently from this growing literature, mainly employing manual coding methods that are difficult to scale to large datasets (Beauchamp, 2020).

We present a systematic review of 67 empirical studies of online dialogue to identify *Textual Indicators of Deliberative Dialogue* (TIDDs). TIDDs are defined as text-based measures of online dialogue quality under a deliberative model (e.g., disagreement, incivility, justifications). TIDDs aim to bridge deliberative theory with the growing empirical literature using online dialogue data. Each TIDD measures a single construct using machine learning, manual coding, or rule-based automatic text analysis. TIDDs are reviewed for their reliability, criterion validity, and construct validity for measuring six deliberative dimensions: rationality, interactivity, equality, civility, constructiveness, and common good reference (Friess & Eilders, 2015).

The review's goal is to identify TIDDs and evaluate their applicability for deliberative theory. TIDDs provide a snapshot of what is measurable in online dialogue and, therefore, a list of text-level variables available to researchers for predicting desirable post-dialogue outcomes. We identify 123 TIDDs, evaluating them as having mid-low reliability, low

criterion validity, and high construct validity for measuring civility and rationality in online dialogues. Our results demonstrate the variety of text-based variables used for studying online dialogue whilst highlighting the limitation of the deliberative model for conceptualizing them.

1. Background

In 2019, for the first time in history, a majority (51%, 4 billion people) of the world's population were using the internet (International Telecommunications Union, 2020). Many internet users are communicating, either publicly through social networking sites, or privately through semi-synchronous "chats" (Yao & Ling, 2020). Social networks are viewed through a normative lens by the social science literature. Optimists view social networks as a place for discussing the world's problems from multiple perspectives and finding consensus on courses of action (e.g., Bohman, 2004). Pessimists view social networks as entrenching existing political binaries through "echo chambers" that undermine any meaningful consensus (e.g., Sunstein, 2018). These opposing views demonstrate the need for understanding how certain dialogue structures lead to desirable (e.g., consensus) and undesirable outcomes (e.g., echo chambers).

Online dialogue produces behavioral trace data (Lampe, 2013). Trace data is unobtrusive, meaning behaviors are observed in naturally occurring contexts (Webb et al., 1966; Wu & Taneja, 2020). Trace data are normally recorded digitally and, therefore, predominantly relate to people behaving on the internet through a computing device (Howison et al., 2011). Using behavioral trace data enables the empirical study of social processes in near real-time (Lampe, 2013), including how and why the observation of certain online dialogues may lead to positive or negative outcomes. This review asks how textual trace data can be operationalized for measuring online deliberation. Specifically, what constructs are currently measured in online dialogue and how applicable are they to deliberative theory.

1.1. Online dialogue and deliberation

A dialogue is defined as a minimum of two people using a semiotic system to communicate about something (Linell, 2017, p. 302). This is conceptualized as a Self-Other-Object relationship, where two or more selves come together to discuss any object of interest. A dialogue, therefore, comprises all Self and Other observable communicative behaviors on one or several discussed topics (Object).

"Online" dialogue is used as shorthand for *public, asynchronous, text-based* dialogues that happen on the internet. These dialogues involve people coming together with strangers to openly discuss a topic. Online dialogues are asynchronous because participants are not required to immediately reply to each other. They are public because most people can freely observe or participate in them. Finally, they are text-based because participants use written language to communicate with each other.

The term "online dialogue" is preferred over the commonly used "computer-mediated communication" (e.g., Atai & Chahkandi, 2012; Chua & Chua, 2017; Di Blasio & Milani, 2008) due to the public connotation. Computer-mediated communications include private messaging, video conferencing, and emails, which are not necessarily public, asynchronous, or text-based. We focus on online dialogues as their quality is relevant to deliberative theorists, who argue that democratic societies are partly sustained by dialogues conducted in the public sphere (Dahlberg, 2001; Friess & Eilders, 2015; Graham & Wright, 2014; Janssen & Kies, 2005; Sunstein, 2018).

Habermas, who conceptualized contemporary deliberative theory, argues that a healthy democracy is maintained by a public sphere where dialogues strive toward an "ideal speech situation" (1981). The ideal speech situation has four principles (Habermas, 2008, p.

50): (1) “publicity and inclusiveness”, nobody should be excluded if they can contribute; (2) “equal rights to engage in communication”, everyone should have the same opportunity to speak; (3) “exclusion of deception and illusion”, participants should mean whatever they say; (4) “absence of coercion”, nobody should try to silence others for their own merit. According to Habermas, when citizens work towards these ideals, their dialogues can generate solutions to societal problems.

Habermas’ ideal speech situation concerns deliberation, “a process where people, often ordinary citizens, engage in reasoned communications on a social or political issue in an attempt to identify solutions to a common problem and to evaluate those solutions” (Stromer-Galley, 2007, p. 3). Deliberation has an “input”, “throughput”, and “output” stage (Friess & Eilders, 2015; adapted from Wessler, 2008). Input refers to the social, cultural, and physical context where a dialogue takes place. Throughput refers to the quality of dialogue as it is procedurally achieved. Finally, output refers to the outcomes of dialogues independent of the process.

Our review focuses only on the throughput stage of deliberation, which we term deliberative dialogue. Friess and Eilders (2015) identify six dimensions of deliberative dialogue (table 1). We use these dimensions to represent the deliberative perspective, as they are the most recent and comprehensive effort to summarize the deliberative qualities of online dialogue (Beauchamp, 2020, p. 329).

Table 1 Friess and Eilders dimensions of dialogue indicating deliberation (2015, p.323)

Dimension of deliberative dialogue	Definition:
Rationality	Refers to the degree of rational and reasoned behaviors evidenced in the dialogue text. These behaviors include the reasoning and logic used by participants in their communication (Friess & Eilders, 2015, p. 328). Rationality often involves claims and justifications made for them, as well as how well participants keep to the topic at hand (Graham & Witschge, 2003; Stromer-Galley, 2007).
Interactivity	Refers to how participants interact with each other to deliberate (Friess & Eilders, 2015). There are "formal" (structural) dimensions of interactivity, such as the number of participants and turns taken, and "substantial" (cognitive) dimensions of interactivity, such as the degree of attention the participants are paying to each other (Trénel, 2004).
Equality	Refers to the "equal opportunity to articulate arguments and to reply to other participants' claims" (Friess & Eilders, 2015, p. 330). This dimension also includes the input stage of deliberation (i.e., the context), but is demonstrated in the text by whether participants interact at relatively equal rates. Equality may also be measured by looking at the distribution of self-reported demographic information (e.g., gender, race, nationality, etc.) available in the text.
Civility	Refers to participants being considerate and polite towards each other. This dimension is also termed "respect", which includes listening (Friess & Eilders, 2015, p. 330).
Common Good Reference	Refers to participants making justifications and arguments that relate to the common good.
Constructiveness	Refers to participants being sincere and productive in their interactions. This is characterized by a degree of intent and execution of problem-solving.

1.2. Dialogue research approaches

Dialogue research can be "descriptive" or "prescriptive" (Stewart & Zediker, 2000). Prescriptive approaches define dialogue in terms of desirable outcomes (Kim & Kim, 2008; Stewart & Zediker, 2000). Prescriptive approaches, including deliberative theory, view dialogue as essential to "growth, development, and positive change" for individuals, communities, and societies at large (Cooper et al., 2013, p. 82). Dialogue, under this view, always produces positive societal outcomes.

Descriptive approaches, in contrast, view dialogue as a "pervasive" feature of human behavior that should be described empirically (Stewart & Zediker, 2000, p. 225). The descriptive approach does not tie specific dialogue structures to ideal outcomes. Instead, it regards dialogue as the primary mechanism for coordinating human social behaviors

(Gergen et al., 2004). Dialogue, under this view, produces many societal outcomes, both positive and negative.

A key problem with prescriptive approaches is that high-quality dialogue is defined independent of context, thereby prescribing what communicative behaviors are desirable regardless of outcomes. Descriptive approaches avoid this problem by studying the diversity of potential outcomes without making any prior normative recommendations on communicative behaviors (Gillespie et al., 2014). A prescriptive approach instead obfuscates the possibility that non-ideal communicative behaviors may lead to desirable outcomes.

Deliberative theory exemplifies the prescriptive approach by arguing that when dialogue is not “fair and equitable”, the outcomes will necessarily be distorted (Cooper et al., 2013, p. 80). Habermas’ ideal speech situation has been criticized for being prescriptive despite his assuming a descriptive approach elsewhere (Kim & Kim, 2008, p. 56). Nonetheless, deliberative theory has remained focused on prescription, developing Habermas’ ideals for deliberation into manual coding frameworks to identify "good" dialogue (e.g., Graham, 2008; Graham & Witschge, 2003; Steenbergen et al., 2003).

Manual coding, however, is impractical for the scale of online dialogue. We agree with Beauchamp (2020, p. 323) that adopting Natural Language Processing (NLP) in deliberative theory would enable more rigorous testing of its conceptual frameworks than is currently done. NLP is concerned with studying organic human communication and the automatic analysis of text (Boyd et al., 2020; Boyd & Schwartz, 2021). This includes extracting information, applying classifications, and measuring the frequency of observable variables (Mehl & Gill, 2010, p. 109).

Automatic measurement provides three improvements over manual coding. First, manual coding takes substantial time and resources to complete, even on small datasets. In

contrast, automated measurement, once developed, is cost-effective and scalable. Second, a manual coding framework may have difficulties replicating when used by new researchers, diminishing the reliability of findings. In opposition, automated measures are perfectly reliable, producing identical results when applied to the same data. Third, automated measurement enables the possibility of real-time monitoring of the quality of online dialogues, which would be impossible with manual coding.

Empirical studies using NLP represent the descriptive approach to studying dialogue quality. When operationalizing online dialogue, NLP studies are not constrained by deliberative theory (see Beauchamp, 2020, p. 331). Instead of using a conceptual framework to derive measures, they can do so by observing dialogues. Thus, identifying measures from a wide empirical literature (including NLP studies) benefits deliberative theory by providing alternative constructs for predicting desired outcomes. To conceptualize and review the diversity of measures from the empirical literature, we introduce the concept of *Textual Indicators of Deliberative Dialogue*.

1.3. Textual Indicators of Deliberative Dialogue

Textual Indicators of Deliberative Dialogue (TIDDs) are text-based measures of online dialogue quality relevant to a deliberative model. TIDDs can be measured either within a turn, between turns, or across the whole dialogue by aggregating turns. TIDDs exclude trace data that occur independently of the dialogue text or is domain-specific, such as likes or click-through rates. TIDDs include structural features of online dialogue (e.g., number of turns, number of replies) that are universal in text-based communication. TIDDs are conceptualized as a bridge between the deliberative (prescriptive) and empirical (descriptive) literatures.

TIDDs can be measured using manual coding, fully automated methods, or supervised machine learning. Manual coding refers to people annotating text data for the presence of target phenomena, formalized as “content analysis” (Krippendorff, 2018). Fully automated methods refer to non-machine learning automatic text analysis, such as dictionary methods. Dictionary methods – such as the Linguistic Inquiry and Word Count tool (LIWC, Pennebaker et al., 2001) – measure a construct in text by counting the occurrence of relevant words (e.g., people’s emotions as indicated by emotional words). Finally, supervised machine learning methods fall between automated and manual coding traditions. They require a manually coded dataset to “learn” the best way to predict a response variable based on hand-coded data. This includes newly trained algorithms for a specific context, or pre-trained algorithms such as Google’s Perspective Application Programming Interface (API, 2021), which identifies uncivil communicative behaviors in text.

1.4. The present study

We systematically review TIDDs for measuring dialogue quality under a deliberative model. Our study is unusual in reviewing constructs from the empirical literature, rather than examining their results. Therefore, we did not use a specific protocol but followed the PRISMA guidelines (Page et al., 2021) where applicable.

The review has three research questions to assess the viability of using TIDDs as indicators of deliberative dialogue. Combined, these three research questions help identify synergies between the deliberative and empirical literatures.

RQ1: What is the reliability of the TIDDs?

Reliability reflects the degree to which results obtained by a measurement process are reproducible (John & Benet-Martínez, 2014, p. 342; Shrout & Lane, 2012, p. 302). This

research question addresses the replicability of a TIDD's measurement method employed by the studies: manual coding, fully automated methods, or supervised machine learning.

RQ2: What is the criterion validity of the TIDDs?

Criterion validity reflects how accurately a measurement (or scale) correlates with a relevant outcome (Bryant, 2000, p. 106). This research question addresses how well TIDDs correlate with outcomes external to the dialogue text.

RQ3: What is the construct validity of the TIDDs?

Construct validity reflects how accurately a variable measures a target concept (Cronbach & Meehl, 1955). This research question addresses how well TIDDs fit the deliberative dialogue model (table 1, Friess & Eilders, 2015). Accordingly, RQ3 addresses the extent to which the identified TIDDs measure rationality, civility, interactivity, constructiveness, equality, and common good reference.

2. Methods

2.1. Search strategy

Our target literature was any empirical article with a systematic operationalization of online dialogues. We targeted studies using NLP but also included those using manual coding. We identified the studies through two searches on three databases – Scopus, PsychInfo, and EmBase – in October 2020. The first search focused on Friess and Eilders (2015) dimensions and the second focused on the quality of dialogue. To further constrain the searches, we developed three additional lists of words. The first identifies studies using online data. The second identifies studies about dialogue. The third identifies empirical

studies with a systematic methodology. The full lists of search terms are in the supplementary materials (A – 1).

For a study to appear in the results, a word from each of the three lists needed to be in the “title, abstract, or keywords” for the Scopus search, or the abstract for the PsychInfo and EmBase searches. We chose the latter search option as it most resembled the Scopus option. All searches were limited to articles published in English with no time-period constraints.

2.2. Inclusion/exclusion criteria

Studies were included if they were empirical, published in a journal or conference proceedings, and used online dialogue data. A study was considered empirical if it reported a clear and systematic method in either the abstract or body of the text. Online dialogue data was defined as public, asynchronous, text-based, and naturally occurring interactions involving two or more individuals.

Studies were excluded if were published as a book chapter, do not include any text-level variables, or examined dialogues that were not online. We, therefore, excluded studies using exclusively private computer-mediated communications (e.g., direct messaging, emails, etc.), dialogues elicited through experimental conditions, or surveys about online dialogues. Detailed inclusion/exclusion criteria are in the supplementary materials (A – 2).

2.3. Data extraction and analysis

For each TIDD, we extracted a definition, the unit of analysis, the measurement methods and associated statistics, and the results of the analyses conducted. Once the TIDDs were identified, we grouped identical (or highly similar) measures under an umbrella term.

For RQ1, the reliability of the TIDDs was estimated through an iterative process outlined in table 2. Each TIDD was classified as being measured using manual coding, supervised machine learning, or fully automated extraction. Manual coding involves subject specialists qualitatively scoring TIDDs in dialogues. Supervised machine learning measures TIDDs using algorithms trained on manually coded data. Automated extraction measures TIDDs using an existing NLP tool to extract a variable computationally (e.g., counting words of a type).

We scored all automated TIDDs reliability as "high" because, when applied to the same dialogue, the TIDD will always produce identical results. We assessed manual coding TIDDs' reliability according to reported interrater reliability statistics (e.g., Krippendorff's α (1970), Cohen's κ (1960), or Scott's pi (Scott, 1955)) and machine learning TIDDs according to their reported accuracy statistics (e.g., Area Under the Curve, F-score). For both, TIDDs with relevant statistics above 0.70 were classed as high reliability, those between 0.50 and 0.70 as mid reliability, and those under 0.50 as low reliability. These cutoff levels were chosen to provide a comparable estimate of reliability across different research traditions.

Table 2 TIDDs Evaluation method

	Evaluation method		
Resulting level	Reliability	Criterion Validity	Construct Validity
High	If manual coding: interrater Reliability statistic > 0.70; if machine learning F-score/AUC > 0.70. If automated.	If TIDD correlates with, or predicts, one or more outcomes outside the dialogue.	If a TIDD has discriminant content validity for a dimension of deliberation and the quality of dialogue.
Mid	If manual coding: interrater Reliability statistic > 0.50, < 0.70; if machine learning: F-score/AUC > 0.50, < 0.70.	If TIDD correlates with, or predicts, at least one outcome outside the dialogue.	If a TIDD has discriminant content validity for multiple dimensions of deliberation and the quality of dialogue.

Low	If manual coding: interrater Reliability statistic < 0.50 or unreported; if machine learning F-score /AUC < 0.50 or unreported.	If a TIDD is not correlated with, or does not predict, any outcome at a statistically significant level.	If a TIDD has discriminant content validity for a dimension of deliberation but not for the quality of dialogue and vice versa.

For RQ2, the criterion validity of the TIDDs was established by examining the studies' results and the variables predicted by TIDDs. The process of determining criterion validity is summarized in table 2. We first noted whether a TIDD correlates with any outcome variables independent of the dialogue text. The strength of these correlations then determines the TIDD's criterion validity rating. We do not consider instances where a TIDD is an outcome (i.e., dependent variable) as demonstrating criterion validity.

For RQ3, the construct validity of the TIDDs is determined by whether they have discriminant content validity (Johnston et al., 2014) for measuring the six dimensions of deliberative dialogue. In big data research, construct validity is difficult to establish as studies normally use naturally occurring behavioral trace data (e.g., online dialogues) instead of survey data (Braun & Kuljanin, 2015; Xu et al., 2020). For standard survey data, researchers typically employ a Confirmatory Factor Analysis (CFA) to estimate the construct validity of their measures (Bryant, 2000). CFA requires a minimum of three measures of a target construct to obtain a model (Anderson & Rubin, 1956). This is easily done with a survey, where new items can be added and tested at will. With naturally occurring trace data, however, behaviors will likely occur at varying frequencies, resulting in lots of missing data for behaviors that are not regularly observed (Braun & Kuljanin, 2015, p. 523). This often renders CFA untenable.

As an alternative to conventional methods of assessing construct validity in big data contexts, the literature recommends using "subject matter experts (SMEs) to rate the relevance of behavioral trace variables or measuring a construct of interest" (Braun & Kuljanin, 2015, p. 525). To make this process more robust, we propose using discriminant

content validity (Johnston et al., 2014) to assess how well the TIDDs discriminate between a set of conceptually relevant dimensions. Eleven coders (of MSc level in psychology or linguistics and including both authors) assigned the TIDD to one of the six dimensions or an "other" category and provided a confidence score. We also had raters assess how well a TIDD measures the quality of dialogue independent of the dimensions. These confidence scores allowed us to test the viability of the deliberative dimensions for conceptualizing the TIDDs.

3. Results

3.1. Descriptive statistics

Figure 1 shows a PRISMA flowchart (Moher et al., 2015) for determining the final list of studies included in the systematic review. The results of the searches were first combined (n=3,908) and any duplicates removed (n=3,185). We then examined the titles and abstracts to determine whether studies should be included (n=208). Additional studies (n=20) were subsequently added manually after being identified in bibliographies as relevant for the review. The remaining studies (n=225) were then assessed for their viability using the entire text. This produced the final list of studies (n=67). Both assessment stages used the same inclusion and exclusion criteria.

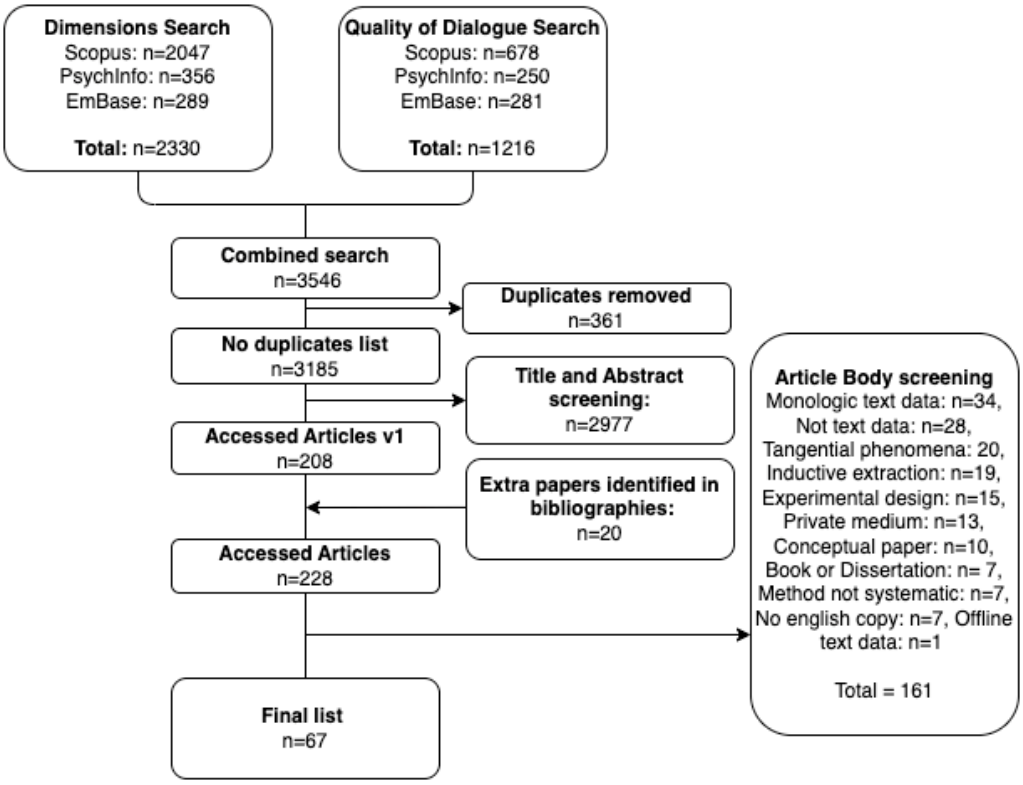

Figure 1 PRISMA flowchart (Moher et al., 2015)

Most studies were published in peer-review journals (n=40, 60%), with the remaining being conference papers (n=25, 37%) or preprint studies on the arXiv database (n=2, 3%). Figure 2 plots when the studies were published and whether they involved a social science, computer science, or a combined approach to deliberative dialogue. We categorized a study based on the journal of publication and the “subject area and category” listing on the ScimagoJR database (SJR, n.d.). Most of the studies were published after 2010 (n=65, 97%). We find an equal number of studies from social science and computer science traditions (n=24, 36%) and 19 studies (28%) from journals taking a mixed approach. Figure 2 shows the sharp increase in interest from both social science and computer science approaches in the 2010-2020 time-period.

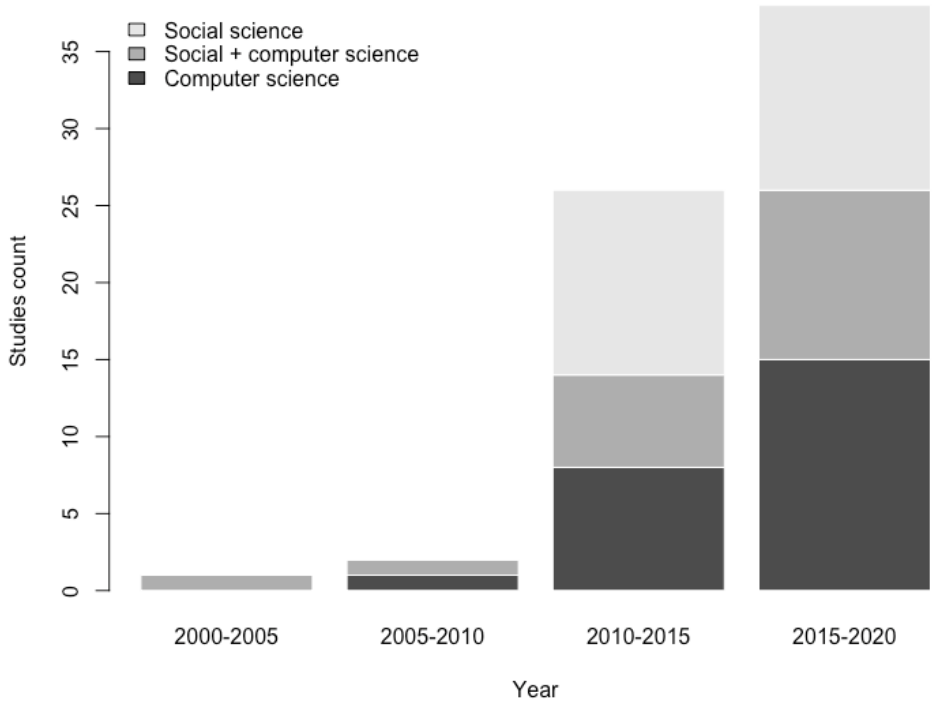

Figure 2 Distribution of empirical studies by year and discipline

In the 67 studies (see supplementary materials A – 5), we initially identified 221 TIDDs. Of these, the majority were measured using manual coding frameworks (n=170, 77%), with 32 (14%) using a machine learning algorithm, and 19 (9%) using a fully automated method.

We then grouped TIDDs with identical or highly similar conceptual definitions. We chose to leave similar measures separated when there was an ambiguous conceptual difference (e.g., personal insult and name-calling). This reduced the TIDDs to 123 independent measures (supplementary materials B & A – 4). Each study is represented in this list at least once. Of the final list of TIDDs, 103 (84%) are measured within a single turn, 12 (10%) are between turns, and eight (6%) use the entire dialogue by aggregating texts together (see table 3).

Table 3 Sample size and types reported (rounded to nearest full number)

Unit of analysis	Mean	Standard Deviation	Median	Min	Max	N %
Turns	1,471,522	8,211,346	3,051	120	60,300,000	57 (89%)
Interactions	9,464	36,804	112	2	215,000	39 (60%)
Participant	14,933	49,356.62	779	11	209,776	18 (28%)

Table 4 shows the ten most frequently measured TIDDs across the 67 studies. Disagreement is the most common, cited by twelve studies, and refers to whether a turn is agreeing or disagreeing with a previous turn in the dialogue. In joint second, both cited ten times, was a general measure of incivility, which measures the degree of civil or uncivil behavior observed in the turn, and justification, which measures whether the claims made by participants in the dialogue are justified.

Table 4 Ten most frequent TIDDs (out of 123)

TIDD	Definition	N (%) of studies	Extraction
Disagreement	The turn content shows disagreement with an Other's position previously given.	12 (18%)	Manual Coding & Machine Learning
Incivility	The turn content shows the Self behaving uncivil manner.	10 (15%)	Manual Coding, Machine Learning, & Automated
Justification	The turn provides a reason for the claims made and a position on these claims.	10 (15%)	Manual Coding & Machine Learning
Position on topic	The turn demonstrates the Self position on a given topic.	6 (9%)	Manual Coding
Claim	The turn makes a claim about a given topic (e.g., a fact).	5 (7%)	Manual Coding & Machine Learning
Linguistic alignment	The degree of alignment between participant's language use (both in semantics and grammar) across the dialogue.	5 (7%)	Manual Coding & Automated
Sentiment	The measurable sentiment of the words used in a turn or across an aggregate of turns (positive, negative, neutral).	5 (7%)	Manual Coding & Automated

Argumentation	The turn is considered argumentative by expressing a clear position on a topic.	5 (7%)	Manual Coding & Automated
Personal insult	The turn contains an insult aimed at an attribute specific to an Other.	5 (7%)	Manual Coding & Machine Learning
Number of responses	The number of turns answering a direct question from an Other.	4 (6%)	Manual Coding

3.2. RQ1 – What is the reliability of the TIDDs?

Table 5 shows the measurement methods and the reliability of the identified TIDDs. We evaluated 54 (44%) TIDDs as low reliability, 37 (30%) TIDDs as mid reliability, four (3%) as mid to high reliability, and 28 (23%) as high reliability.

Table 5 Reliability of TIDDs and measurement method

	Reliability				
Measurement Method	Low	Mid	Mid-High	high	Total:
Manual Coding	54 (44%)	26 (21%)		20 (16%)	100 (81%)
Manual Coding & Machine Learning		6 (5%)	1 (1%)		7 (5%)
Manual Coding & Automated		2 (2%)	3 (2%)		5 (4%)
Manual Coding, Machine Learning & Automated		1 (1%)			1 (1%)
Machine Learning		2 (2%)		2 (2%)	4 (4%)
Machine Learning & Automated				1 (1%)	1 (1%)
Automated				5 (4%)	5 (4%)
Total:	54 (44%)	37 (30%)	4 (3%)	28 (23%)	123 (100%)

A majority of the TIDDs (n=100, 81%) were measured using exclusively manual coding. This includes 54 low reliability TIDDs, which either had no reported interrater reliability statistic or a statistic of value below 0.50. We found that four (3%) TIDDs were

measured exclusively with machine learning and five (4%) using exclusive automated measures. The remaining 14 TIDDs were measured using multiple measurement methods.

Interrater reliability statistics were identified for 61 TIDDs with 121 associated values. These interrater reliabilities range from 0.32 to 1.00, have a mean of 0.78 and a standard deviation of 0.12. Of the 121 reported values, 72 (59%) report Krippendorff's α, 30 (25%) report Cohen's κ, and the remaining 19 (16%) use an alternative statistic such as Maxwell's RE, Scott's Pi, or the Pearson product-moment correlation. Machine learning accuracy statistics were identified for 13 TIDDs with 29 reported values. The F-score was the most reported statistic (n=22, 76%), ranging from 0.49 to 0.91, with a mean of 0.73 and a standard deviation of 0.12.

3.3. RQ2 – What is the criterion validity of the TIDDs?

We found seven studies (Augustine & King, 2019; Coe et al., 2014; Eschmann et al., 2021; Han, 2018; Hopp et al., 2020; Loveland & Popescu, 2011; Zhang et al., 2013) reporting concurrent validity (i.e. correlations between TIDDs and a co-present outcome) for 12 TIDDs (see table 6). A majority of TIDDs (n=111, 90%) were rated as having low criterion validity because no correlations with outcomes beyond the dialogue were given. We evaluated three (2%) TIDDs with high criterion validity and eight (7%) with mid criterion validity.

Table 6 High and Mid rated criterion validity TIDDs

TIDD	Criterion Validity	Outcome variables summary
Disagreement	High	Coherence (Pearson's r = -0.27) - (Augustine & King, 2019); number of posts and contributors ratio (β = 1.952, $p < 0.01$)[a] - (Loveland & Popescu, 2011)
Incivility	High	Incivility self-report measures (average over several Kendall's t = 0.34, $p <= 0.05$) - (Hopp et al., 2020)

Conceding	High	Number of tweets (β = 0.045, p = 0.049)[b], number of replies (β = 0.018, p = 0.023)[b] - (Eschmann et al., 2021)
Justification	Mid	Moderation affordances of forum (z = 10.99, p <0.001)[a] (Zhang et al., 2013)
Argumentation	Mid	Number of posts and contributors ratio (β = 0.887, p < 0.01)[a] - (Loveland & Popescu, 2011)
Number of function words	Mid	Perceived credibility (β = 0.153, p < 0.05)[a], perceived intent (β = 0.155, p < 0.05)[a] - (Han, 2018)
Impolite words	Mid	Moderation affordances of forum (z = 7.63, p < 0.001)[a] - (Zhang et al., 2013)
Name-calling	Mid	Thumbs down (t-test between civil and uncivil groups and Number of thumbs down: t = 3.74, p < 0.001) - (Coe et al., 2014)
Personal insult	Mid	Thumbs down (t-test between civil and uncivil groups and Number of thumbs down: t= 3.74, p < 0.001) - (Coe et al., 2014)
Sentiment	Mid	Perceived credibility (β = 0.073, p < 0.05)[a], perceived attraction (β = 0.069, p < 0.05)[a], perceived intent (β = 0.065, p < 0.05) [a] - (Han, 2018)
Turn similarity	Mid	Perceived credibility (β = -0.17, p < 0.05) [a], perceived attraction (β = -0.146, p < 0.05) [a], perceived Competence (β = -0.152, p < 0.05) [a], perceived intent (β = -0.166, p < 0.05)[a] - (Han, 2018)
Vulgarity	Mid	Thumbs down (t-test between civil and uncivil groups and Number of thumbs down: T = 3.74, p < 0.001) - (Coe et al., 2014)

[a]Linear regression, [b]Logistic regression

Three studies are notable for their consideration of criterion validity in their research designs. First, Han (2018) had individuals rate the quality of Twitter dialogues through a survey along four dimensions (Credibility, Attraction, Intent, and Competence) and proceeded to explore whether TIDDs measured in the text predict the results. Despite their results showing modest associations between the TIDDs measured and perceived quality variables, this method of evaluating criterion validity appears robust.

Second, Hopp and colleagues (2020) find positive and statistically significant correlations between self-report and text-based measures of incivility. The authors use Google's Perspective API to automatically identify uncivil political communication in Facebook and Twitter data, collected from participants who reported on how frequently they engaged in uncivil dialogue.

Finally, Loveland and Popescu (2011) create a “quality of deliberation” metric which involved a ratio between the number of contributions and contributors (see p. 693). They then use this metric to explore whether specific TIDDs (e.g., disagreement and argumentation) predict this ratio. Whilst it is useful to see these correlations, the reason why the ratio represents “quality” is unclear.

3.4 RQ3 – What is the construct validity of the TIDDs?

Across all the TIDDs, the eleven coders had a Krippendorff’s α (1970) of 0.39 and 15% agreement. This indicates that many TIDDs were allocated to more than one dimension. Despite this, we still found the majority of TIDDs (n=77, 62%) to have high construct validity in measuring a single dimension and the quality of online dialogue. Only five (4%) TIDDs were evaluated as having mid construct validity and the remaining had low construct validity (41, 33%).

Table 7 shows that civility has 41 (33%) TIDDs, rationality has 33 (27%) TIDDs, and equality has three (2%) TIDDs with statistically significant discriminant content validity for both the dimension and quality of dialogue. All of these were rated as high construct validity. There are 41 TIDDs with no statistically significant content validity for a dimension but do for the quality of dialogue and one TIDD which is neither significant for a dimension nor the quality of dialogue (group decision).

Table 7 Discriminant content validity results

	Discriminant Content Validity results	
Construct	Exclusive to dimension	Shared between dimensions

Rationality	Additional knowledge, affirmation, argumentation, argumentative function, ask for detail, ask for evidence, asking for a yes or no, balanced view, citing, challenge, claim, clarification request, counterclaim (polite), demonstrating understanding, disagreement, elaboration (context), exclamation, genuine questions, social support with information, justification (external), linguistic alignment, lying, moral values given, narrative claim, object function, topic relevance, originality, position on topic, provide summary of dialogue, providing context, providing information, supplication, topic as object	Conceding, ideology, irrelevancy claim, judgment of ideological positions, turn similarity
Civility	Abusive language, accusation, accusation of incompetence, aggressive emotions, aggressive language, anti-white prejudice, apology, avoiding argument, color-blind racism, consider other's opinions, criticizing other's talk, disclosing feelings, dismissive tone, disrespecting norms, norm-rejecting, gratitude, hate speech, impolite words, incivility, interjection, intolerance of incivility, irony, metatalk, name-calling, nastiness, number of turns, offensive remark, overt racism, personal insult, politeness, prosocial behavior, provocative/extremist statements, respect, responsiveness, sarcasm, sentiment, tone, divisive topic, uncivil language, use of all caps, warning an other	Conceding, irrelevancy claim, judgment of ideological positions, turn similarity
Equality	Authority signaling, demographic information, diversity of perspectives	Ideology
Indicative of quality with no clear dimension	Addressivity, asking questions, communication style (broadcasting vs engaging), community appreciation, concern for others, constructive contribution, contradiction, disclosing thoughts, elaboration, elaboration (example), elaboration (explanation), engagement, engaging others' experiences, number of function words, important words, information exchange, informative richness, interactive, perspective taking, informative richness, interpersonal relationship as topic of dialogue, justification, mild scolding, turn timing, number of responses, number of words used, propose alternative solution, propose solution, reciprocation of self-disclosure, dialogue regulation, repair strategies, request in turn, requesting, turn similarity to target topic, social support, stereotyping, making suggestions, topics under discussion, toxic comments, weighing pros and cons, vulgarity	

All $p < 0.05$ for dimensions (rationality, civility & equality) and quality with Wilcoxon (one-sided) sample-rank test

Table 8 shows the 20 best performing TIDDs and their reliability, criterion validity, and construct validity. A TIDD was included if it had either no "low" ratings or at least two "high" ratings. We found only one TIDD (disagreement) is rated high on all three evaluations. One TIDD (incivility) has high criterion validity, high construct validity, and mid reliability. One TIDD (sentiment) has high construct validity, mid criterion validity, and mid reliability. The remaining 16 TIDDs have high construct validity and reliability, but low criterion validity.

Table 8 Best performing TIDDs across reliability, criterion validity, and construct validity ratings.

TIDD	Definition	Construct validity	Criterion validity	Reliability
Disagreement	The turn content shows disagreement with an Other's position previously given.	High	High	High
Incivility	The turn content shows the Self behaving in an uncivil manner.	High	High	Mid
Sentiment	The measurable sentiment of the words used in a turn or across an aggregate of turns (positive, negative, neutral).	High	Mid	Mid
Accusation	The turn accuses an Other in the dialogue of doing something (within or beyond the dialogue).	High	Low	High
Ask for evidence	The turn contains a question requesting or demanding evidence of an Other.	High	Low	High
Avoiding argument	The turn demonstrates an explicit attempt to avoid an argument (e.g., by stating "I don't want to argue").	High	Low	High
Citing	The turn contains a citation of an Other in the dialogue (e.g., using quotation marks or by rephrasing).	High	Low	High
Consider other's opinions	The turn takes into account an Other's feelings in the comments and responses (e.g., by providing "trigger warnings").	High	Low	High
Demographic information	The turn contains a reference to the demographics of the participants (e.g., gender, ethnicity, nationality, etc.).	High	Low	High
Hate speech	The turn contains hateful speech (e.g., racial slurs, derogatory comments towards a group, homophobia, etc.).	High	Low	High
Impolite words	The number of impolite words present in a turn or across an aggregate of turns?	High	Low	High
Interjection	The turn is interjecting the current path of a dialogue (e.g., replying on Twitter when a Self has indicated they need more turns).	High	Low	High
Intolerance of incivility	The turn shows intolerance for others' incivility.	High	Low	High
Justification (external evidence)	The turn is justifying a claim made in a current or prior turn using external evidence (e.g., links to another website).	High	Low	High
Linguistic alignment	The degree of alignment between participant's language use (both in semantics and grammar) across the dialogue.	High	Low	High
Lying	The turn content is observably insincere and/or deceitful.	High	Low	High
Responsiveness	The turn is responding to another when expected (e.g., answering questions, reciprocating greetings, etc.).	High	Low	High
Social support with information	The turn provides information to an Other that is supportive.	High	Low	High
Supplication	The turn uses religious language to justify claims.	High	Low	High
Uncivil language	The turn contains language assessed to be uncivil (e.g., being vulgar, treating serious topics with humor, etc.).	High	Low	High

Table 9 shows the top discrimination failures in TIDDs of low construct validity. This equates to pairs of dimensions that appeared most frequently when aggregating coders' categorization of TIDDs. We observe considerable crossover between all dimensions. Constructiveness displays the most discrimination failures, appearing in the top three pairs with interactivity, civility, and common good reference.

Table 9 Discrimination failures between dimension pairs for low construct validity TIDDs (full version in supplementary materials A – 3).

Dimensions appearing together	Count (percentage of total TIDDs)
Constructiveness + Interactivity	21 (17%)
Constructiveness + Rationality	10 (8%)
Constructiveness + Common good reference	7 (6%)
Interactivity + Equality	7 (6%)
Interactivity + Civility	7 (6%)
Constructiveness + Civility	5 (4%)
Constructiveness + Equality	5 (4%)
Interactivity + Rationality	5 (4%)
Interactivity + Common good reference	5 (4%)
Civility + Equality	4 (3%)
Civility + Common good reference	3 (2%)
Common good reference + Equality	3 (2%)
Common good reference + Rationality	1 (1%)

4. Discussion

This review introduces TIDDs and assesses the viability of using automated text analysis to assess the quality of online deliberative dialogue. The review employs a novel use of Discriminant Content Validity (Johnston et al., 2014) to estimate the construct validity of behavioral trace measures. Discriminant content validity was designed to test the face validity of survey items before pretesting their statistical construct validity, however, it is ideal for testing the construct validity of measures designed for behavioral trace data.

We identified 123 TIDDs from 67 studies and found that, on average, they have weak to medium reliability (RQ1), low criterion validity (RQ2), and high construct validity in measuring civility and rationality (RQ3). These findings reflect the viability of using textual measures to assess the quality of deliberation, and usefulness of deliberative theory to conceptualize textual measures.

Examining the reliability of the TIDDs (RQ1), we find that the empirical literature studying online dialogue is primarily using manual coding to measure constructs. Automatic extraction – both by rule-based (e.g., dictionary methods) or machine learning algorithms – are currently not the norm. The use of preexisting manually coded datasets was common in studies focused exclusively on training a machine learning classifier. Whilst preexisting datasets are useful for model design, they can have validity problems because a machine learning classifier will replicate any biases present in the original data. This limitation is demonstrated by Hoffman, McDonald, and Zachry's (2017) attempt to validate Danescu-Niculescu-Mizil and colleagues' politeness classifier (2013). In the study, the tool does not perform as expected, failing to classify instances of politeness in contexts that were dissimilar to the initial training data. They conclude that future machine learning algorithms will likely improve the prediction of politeness, but current models are limited (Hoffman et al., 2017, p. 12).

The overall reliability of the indicators can be improved in three ways. First, by developing new standardized tools for measuring TIDDs. Any automated tool (such as LIWC and the Perspective API) has high reliability as it produces identical results when used repeatedly on a document. Second, the validity of existing tools should be continuously tested on unseen data to confirm the continued accuracy of results. This could be achieved by checking a machine learning algorithm's predictions against human coders (Hoffman et al., 2017) or focusing on the construct validity of linguistic features extracted before training (see Sao Pedro et al., 2012). Third, researchers should work collaboratively to create and

maintain open source databases of online dialogue, manually coded with relevant constructs. This would help train supervised machine learning algorithms, which would facilitate automation in future analyses.

Examining the criterion validity of the TIDDs (RQ2), we identify more instances of TIDDs being used as dependent variables than independent. This demonstrates a premature assumption in the empirical literature that certain TIDDs are definitive indicators of an online dialogue's quality. This shows how both the deliberative and empirical literatures need to test widespread assumptions about online dialogue outcomes. For instance, incivility can be used to combat oppressive conditions or express familiarity in a community. Incivility, therefore, does not necessarily correlate with undesirable outcomes and low-quality dialogue in all contexts (Chen et al., 2019).

Another example of assuming an outcome is Loveland and Popescu's (2011) "deliberative quality" metric to measure the effects of civility TIDDs on online dialogue. This metric is a ratio of the number of posts divided by the number of contributors to the thread and multiplied by the "proportion of thread posts which responded to a prior post" (p. 695). Whilst civility TIDDs may correlate with this metric in these contexts, the assumption that the metric represents the quality of dialogue is unverified.

We propose three methods for future studies to determine whether TIDDs can be used as outcome measures. First, researchers could employ the method employed in Han's (2018) study. Participants first rank and rate naturally occurring dialogues for their perceived quality. TIDDs are then extracted from the dialogues to test whether they predict the ratings. Second, participants take a survey before and after taking part in online dialogue, rating their attitudes at each stage. This would allow testing of whether attitudes correlate with target TIDDs. Finally, researchers may choose to correlate TIDDs with self-report measures of the same constructs (as done by Hopp et al, 2020).

Examining the construct validity of the TIDDs (RQ3), we found they could only be reliably classified into three of the six Friess and Eilders (2015) dimensions of deliberative dialogue. Of these three dimensions, rationality and civility are the best represented, associated with 74 (60%) TIDDs in total. In contrast, equality is only associated with three (2%) TIDDs. There were also a high number of TIDDs with good discriminant content validity for measuring a broad dialogue quality category but did not measure any of the deliberative dimensions (n=40, 33%).

This suggests the deliberative model is limited for conceptualizing the current ways online dialogues are operationalized by the broader empirical literature. This is likely a result of conceptual crossovers between dimensions. For instance, the opposite of civility and constructiveness appears to be antisocial behaviors for both dimensions. Civility is about people treating each other with respect and constructiveness is about working towards constructive shared outcomes. Therefore, being antisocial appears to be both uncivil and unconstructive.

Our results evidence the need for the prescriptive deliberative literature to adapt their model according to the variety of TIDDs present in the wider empirical literature. Overall, the deliberative model, derived from coherent but abstract principles, does not operationalize parsimoniously when used to analyze actual dialogue. Therefore, we recommend altering the model using principles from descriptive approaches to dialogue (Stewart & Zediker, 2000).

A descriptive approach better represents the current state of the empirical literature and emphasizes the communicative behaviors participants employ toward each other. This model is implicit in Détienne and colleagues' (2016) study, where they focus on the "dialogic functions" of turns in Wikipedia conversations. This involves describing each turn in terms of

what it is doing for the participants and their social-cultural context. We recommend developing a model using the axiomatic definition of dialogue as a Self and Other communicating on an Object. Each TIDD indicates a combination of Self-Other-Object components in a tripartite model: object-focused, other-focused, and intersubjective-focused.

Object-focused TIDDs concern a Self providing information or attitude on the Object of conversation. Prototypical Object-focused TIDDs include conceding, counterclaim, claim, justification, diversity of perspectives, position on topic, and balanced view. These TIDDs concern the justifications participants are making about the topics discussed, and any other TIDDs directed at the exchange of information.

Other-focused TIDDs concern the Self's behaviors towards Others in the dialogue. Prototypical other-focused TIDDs include incivility, judgment of ideological positions, accusation of incompetence, and criticizing others' talk. These TIDDs concern any instance of other-directed communicative behaviors produced by a single individual.

Intersubjective-focused TIDDs concern the overall relationship between Self and Others as they interact over an Object. Prototypical intersubjective-focused TIDDs include disagreement, metatalk, social support, and reciprocation of self-disclosure. These TIDDs pertain to meta dialogue, that is, dialogue that coordinates the perspectives of Self and Other vis-à-vis the Object.

5. Conclusion

This study introduces TIDDs to conceptualize the growing need to measure the deliberative quality of online dialogues. The review provides a practical and theoretical contribution. At a practical level, the TIDDs (supplementary materials B & A – 4) can be applied across a variety of datasets as they are context-independent. They may therefore facilitate automatic comparison of dialogue quality across social media platforms. If

embraced and developed, TIDDs might enable more multidisciplinary research using large-scale online dialogue datasets.

At a theoretical level, we suggest deliberative theory conceptualize TIDDs using a model of Self-Other-Object interactions instead of ideal speech situations. Based on a descriptive definition of dialogue, a Self-Other-Object model better reflects empirical reality than abstract theoretical constructs such as rationality and civility. Adopting this model can help broaden current empirical measures used for predicting desirable outcomes of deliberation and, in turn, how online dialogues can be improved for maintaining democracies.

This systematic review has three important limitations. First, the review did not follow a prospective registration before being conducted. Second, only three databases were used for identifying relevant literature. Future studies should seek to expand on the current TIDDs by targeting databases that were not included in the review (e.g., Web of Science, Google Scholar). Third, because the review focused on measured constructs, we only performed a minimal quality check of studies' analysis methods and results during the screening process. Future reviews may expand on our review by assessing how well TIDDs are analyzed in the target studies. This would provide additional insight into the criterion validity of the TIDDs and, therefore, how they might robustly predict dialogue outcomes.

Computers have become integral to the functioning of societies by enabling previously unimaginable dialogues. With these new dialogues come many unanswered questions about the role of communication in democracies. We have shown how the deliberative literature, which conceptualized dialogue quality before widespread computation, benefits from adapting methods and findings from the wider empirical literature. The TIDDs provide a step towards reliable and valid measures of deliberative

dialogue. With refinement, TIDDs have the potential to help monitor and improve the quality of online dialogue and, as a result, the future of global democracy.

References

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (Vol. 5, pp. 111–150). University of California Press.

Atai, M. R., & Chahkandi, F. (2012). Democracy in computer-mediated communication: Gender, communicative style, and amount of participation in professional listservs. *Computers in Human Behavior*, *28*(3), 881–888. https://doi.org/10.1016/j.chb.2011.12.007

Augustine, G., & King, B. G. (2019). Worthy of debate: Discursive coherence and agreement in the formation of the field of sustainability in higher education. *Socio-Economic Review*, *17*(1), 135–165. https://doi.org/10.1093/ser/mwz020

Beauchamp, N. (2020). Modeling and measuring deliberation online. In B. Foucault Welles & S. González-Bailón (Eds.), *The Oxford handbook of networked communication* (pp. 320–349). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190460518.013.23

Bohman, J. (2004). Expanding dialogue: The internet, the public sphere and prospects for transnational democracy. *The Sociological Review*, *52*(1_suppl), 131–155. https://doi.org/10.1111/j.1467-954X.2004.00477.x

Boyd, R. L., Pasca, P., & Lanning, K. (2020). The personality panorama: Conceptualizing personality through big behavioural data. *European Journal of Personality*, *34*(5), 599–612. https://doi.org/10.1002/per.2254

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Braun, M. T., & Kuljanin, G. (2015). Big data and the challenge of construct validity. *Industrial and Organizational Psychology*, *8*(4), 521–527. https://doi.org/10.1017/iop.2015.77

Bryant, F. B. (2000). Chapter 4: Assessing the validity of measurement. In *Reading and understanding MORE multivariate statistics* (pp. 99–146). American Psychological Association.

Chen, G. M., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media + Society*, *5*(3), 1–5. https://doi.org/10.1177/2056305119862641

Chua, Y. P., & Chua, Y. P. (2017). Do computer-mediated communication skill, knowledge and motivation mediate the relationships between personality traits and attitude toward Facebook? *Computers in Human Behavior*, *70*, 51–59. https://doi.org/10.1016/j.chb.2016.12.034

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cooper, M., Chak, A., Cornish, F., & Gillespie, A. (2013). Dialogue: Bridging personal, community, and social transformation. *Journal of Humanistic Psychology*, *53*(1), 70–93. https://doi.org/10.1177/0022167812447298

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Dahlberg, L. (2001). The internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, Communication & Society*, *4*(4), 615–633. https://doi.org/10.1080/13691180110097030

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 250–259. https://www.aclweb.org/anthology/P13-1025

Détienne, F., Baker, M., Fréard, D., Barcellini, F., Denis, A., & Quignard, M. (2016). The descent of Pluto: Interactive dynamics, specialisation and reciprocity of roles in a Wikipedia debate. *International Journal of Human-Computer Studies*, *86*, 11–31. https://doi.org/10.1016/j.ijhcs.2015.09.002

Di Blasio, P., & Milani, L. (2008). Computer-mediated communication and persuasion: Peripheral vs. central route to opinion shift. *Computers in Human Behavior*, *24*(3), 798–815. https://doi.org/10.1016/j.chb.2007.02.011

Eschmann, R., Groshek, J., Li, S., Toraif, N., & Thompson, J. G. (2021). Bigger than sports: Identity politics, Colin Kaepernick, and concession making in #boycottnike. *Computers in Human Behavior*, *114*, 106583. https://doi.org/10.1016/j.chb.2020.106583

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, *7*(3), 319–339. https://doi.org/10.1002/poi3.95

Gergen, K. J., Gergen, M. M., & Barrett, F. J. (2004). Dialogue: Life and death of the organization. In *The SAGE handbook of organizational discourse* (pp. 39–60). SAGE Publications Ltd. https://doi.org/10.4135/9781848608122

Gillespie, A., Reader, T., Cornish, F., & Campbell, C. (2014). Beyond ideal speech situations: Adapting to communication asymmetries in health care. *Journal of Health Psychology*, *19*(1), 72–78. https://doi.org/10.1177/1359105313500251

Google. (2021). *Perspective application programming interface*. https://www.perspectiveapi.com/#/home

Graham, T. (2008). Needles in a haystack. *Javnost - The Public*, *15*(2), 17–36. https://doi.org/10.1080/13183222.2008.11008968

Graham, T., & Witschge, T. (2003). In search of online deliberation: Towards a new method for examining the quality of online discussions. *Communications*, *28*(2), 173–204. https://doi.org/10.1515/comm.2003.012

Graham, T., & Wright, S. (2014). Discursive equality and everyday talk online: The impact of "superparticipants". *Journal of Computer-Mediated Communication*, *19*(3), 625–642. https://doi.org/10.1111/jcc4.12016

Habermas, J. (1981). *The theory of communicative action*. Cambridge: Polity.

Habermas, J. (2008). *Between naturalism and religion: Philosophical essays*. Cambridge: Polity.

Han, K. (2018). How do you perceive this author? Understanding and modeling authors' communication quality in social media. *PLOS ONE*, *13*(2), e0192061. https://doi.org/10.1371/journal.pone.0192061

Hoffman, E. R., McDonald, D. W., & Zachry, M. (2017). Evaluating a computational approach to labeling politeness: Challenges for the application of machine classification to social computing data. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 52:1-52:14. https://doi.org/10.1145/3134687

Hopp, T., Vargo, C. J., Dixon, L., & Thain, N. (2020). Correlating self-report and trace data measures of incivility: A proof of concept. *Social Science Computer Review*, *38*(5), 584–599. https://doi.org/10.1177/0894439318814241

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12). https://doi.org/10.17705/1jais.00282

International Telecommunications Union. (2020). *Percentage of individuals using the internet*. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/

Janssen, D., & Kies, R. (2005). Online forums and deliberative democracy. *Acta Politica*, *40*(3), 317–335. https://doi.org/10.1057/palgrave.ap.5500115

John, O. P., & Benet-Martínez, V. (2014). Measurement: Reliability, construct validation, and scale construction. In *Handbook of research methods in social and personality psychology, 2nd ed* (pp. 473–503). Cambridge University Press.

Johnston, M., Dixon, D., Hart, J., Glidewell, L., Schröder, C., & Pollard, B. (2014). Discriminant content validity: A quantitative methodology for assessing content of theory-based measures, with illustrative applications. *British Journal of Health Psychology*, *19*(2), 240–257. https://doi.org/10.1111/bjhp.12095

Kim, J., & Kim, E. J. (2008). Theorizing dialogic deliberation: Everyday political talk as communicative action and dialogue. *Communication Theory*, *18*(1), 51–70. https://doi.org/10.1111/j.1468-2885.2007.00313.x

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70. https://doi.org/10.1177/001316447003000105

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. SAGE Publications.

Lampe, C. (2013). Behavioral trace data for analysing online communities. In S. Price, C. Jewitt, & B. Brown (Eds.), *The SAGE handbook of digital technology research* (pp. 236–249). SAGE.

Linell, P. (2017). Dialogue, dialogicality and interactivity: A conceptually bewildering field? *Language and Dialogue*, *7*(3), 301–335. https://doi.org/10.1075/ld.7.3.01lin

Loveland, M. T., & Popescu, D. (2011). Democracy on the web. *Information, Communication & Society*, *14*(5), 684–703. https://doi.org/10.1080/1369118X.2010.521844

Mehl, M. R., & Gill, A. J. (2010). Automatic text analysis. In *Advanced methods for conducting online behavioral research* (pp. 109–127). American Psychological Association. https://doi.org/10.1037/12076-008

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n160

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC. *Mahway: Lawrence Erlbaum Associates*, *71*(2001).

Sao Pedro, M. A., Baker, R. S. J. d., & Gobert, J. D. (2012). Improving construct validity yields better models of systematic inquiry, even with less information. In J. Masthoff,

B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *User Modeling, Adaptation, and Personalization* (pp. 249–260). Springer. https://doi.org/10.1007/978-3-642-31454-4_21

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321–325. https://doi.org/10.1086/266577

Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In *Handbook of research methods for studying daily life* (pp. 302–320). The Guilford Press.

SJR. (n.d.). *Scimago journal & country rank*. SCImago Journal & Country Rank [Portal]. Retrieved 4 September 2020, from https://www.scimagojr.com/

Steenbergen, M. R., Bächtiger, A., Spörndli, M., & Steiner, J. (2003). Measuring political deliberation: A discourse quality index. *Comparative European Politics*, *1*(1), 21–48. https://doi.org/10.1057/palgrave.cep.6110002

Stewart, J., & Zediker, K. (2000). Dialogue as tensional, ethical practice. *Southern Communication Journal*, *65*(2–3), 224–242. https://doi.org/10.1080/10417940009373169

Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. *Journal of Deliberative Democracy*, *3*(1), Article 1. https://doi.org/10.16997/jdd.50

Sunstein, C. R. (2018). *#republic: Divided democracy in the age of social media*. Princeton University Press.

Trénel, M. (2004). *Measuring the deliberativeness of online discussions. Coding scheme 2.2* [Unpublished paper].

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences* (pp. xii, 225). Rand Mcnally.

Wessler, H. (2008). Investigating deliberativeness comparatively. *Political Communication*, *25*(1), 1–22. https://doi.org/10.1080/10584600701807752

Wu, A. X., & Taneja, H. (2020). Platform enclosure of human behavior and its measurement: Using behavioral trace data against platform episteme. *New Media & Society*, 1461444820933547. https://doi.org/10.1177/1461444820933547

Xu, H., Zhang, N., & Zhou, L. (2020). Validity concerns in research using organic data. *Journal of Management*, *46*(7), 1257–1274. https://doi.org/10.1177/0149206319862027

Yao, M. Z., & Ling, R. (2020). "what is computer-mediated communication?"—An introduction to the special issue. *Journal of Computer-Mediated Communication*, *25*(1), 4–8. https://doi.org/10.1093/jcmc/zmz027

Zhang, W., Cao, X., & Tran, M. N. (2013). The structural features and the deliberative quality of online discussions. *Telematics and Informatics*, *30*(2), 74–86. https://doi.org/10.1016/j.tele.2012.06.001