Automated and Hand-Coded Measurement of Deliberative Quality in Online Policy Discussions

Peter Muhlberger
Texas Tech University
College of Mass Communications
University and Broadway MS3082
Lubbock, TX 79409
001-806-791-7723
pmuhl830@gmail.com

Jennifer Stromer-Galley
University at Albany, SUNY
Dept. of Communication, SS340
Albany, NY 12222
001-518-442-4873
jstromer@albany.edu

ABSTRACT

A number of research projects are now underway to assess how IT tools and deliberative techniques could enhance public input into government at all levels. The success of such projects depends to an important degree on objective measurement of the quality of discussion under different utilizations of tools and techniques. This paper explores the value of two techniques for determining the quality of policy discussion in the context of federal agency rulemaking: a human-coded technique and a technique involving statistical bootstrapping of natural language processing data. The hand-coded technique utilizes counts of verbal behaviors that are indicative of intellectual engagement. This includes such behaviors as raising questions, disagreeing, and introducing new topics. The automated technique develops a measure of sophistication of reasoning based on significant co-occurrence of concepts in participants' speech. Such co-occurrence clarifies the network of conceptual relations utilized by participants. The more connections participants make between policy-relevant concepts, the more sophisticated their speech should be. Data consist of discussion of the federal rulemaking issue of network neutrality regulation by a sample of 53 volunteers. Correlations and ordinary least squares regression find that the conceptual connection and hand-coded measures, despite being qualitatively quite different, significantly predict each other and several other measures of sophistication, including two indicators of network neutrality knowledge and sophistication of views of government.

Categories and Subject Descriptors

K.4.0 [**Computers and Society: General**]

General Terms

Measurement, Experimentation, Human Factors, Standardization.

Keywords

Deliberation, Public Engagement, Rulemaking, Measurement, Sophistication, Political Discussion, Natural Language Processing.

1.INTRODUCTION

The issues confronted by government are too often complex, intertwined, and politically intractable. Various interest groups present discordant assessments of the facts and value frames of such national issues as global warming or "wicked planning problems" such as local transportation project choices [11]. To the extent that a majority demands resolution to such problems as global warming, questions remain about their willingness to bear the costs of a realistic program to address the problem [15]. The public often wants government programs, so long as it does not have to pay for them [19], a conundrum that makes policy problems difficult to address.

Given the difficulty of addressing intractable issues and the increasingly serious nature of the problems they represent, government appears to require better engagement with the public —engagement that will educate the public, insert intelligent public input into government decisions, increase consensus, and legitimize policy decisions. Theorists, researchers, and advocates of deliberative democracy propose that deliberative public engagement with government will meet these needs [2,7,8,10]. In another corner, information technology (IT) proponents anticipate that information tools are creating a "networked public sphere" with a more informed, interactive, and engaged public [1,4,13,17].

Researchers have begun to explore the promising possibilities inherent in uniting IT tools and audiences with deliberative methods, combining the power of both IT and deliberation to inform, motivate, and coordinate the public. For instance, the Deliberative E-Rulemaking (DeER) project, from which this paper emerges, is seeking to test the value of two deliberation techniques and such natural language processing tools as question answering and summarization for participants in online discussions of policy issues involved in federal agency rulemakings. The measurement of deliberative quality is important, yet difficult, for such projects.

1.1.The Deliberative Quality Measurement Problem

A difficult hurdle for deliberation and IT projects is determining which tools and deliberative techniques positively influence outcomes for government decisions. A key proposition is that certain IT tools and techniques improve the quality of public inputs into governmental processes. The DeER project experimentally exposes some participants to particular tools and deliberation techniques, while others are exposed to one or the other or neither. But how can researchers know that the quality of public comments coming out of a given condition are better than those from another condition? Without an objective measure of quality, the value of particular tools and techniques cannot be scientifically ascertained and any conclusions drawn are apt to be subjective.

A number of alternatives exist for indirectly determining the quality of a decision, but a direct measure remains illusive. Researchers might, for example, develop a multiple-choice quiz of substantive knowledge regarding the policy under discussion. An increase in knowledge from pre- to post- discussion could be taken as evidence of improved quality of choice [12]. Similarly, researchers may take a change in opinions as evidence of greater sophistication in subsequent choices [12]. Opinions, however, can change due to social pressure or acceptance of biased information [16]. It is also far from clear that deliberation rather than exposure to information materials during deliberative experiments increase multiple-choice factual knowledge [14].

Even if deliberation does not change factual knowledge, proponents will claim that it helps people develop more sophisticated understandings of an issue—for example, by increasing the organization of core concepts. The notion that the organization of concepts, not the mere possession of factual bits, matters to the sophistication of reasoning is prevalent in psychology [21]. This suggests that the use of factual knowledge as an indicator of sophistication of reasoning is at best an indirect approach to measurement.

The measurement of the sophistication of policy recommendations coming out of public deliberations may be more directly achieved by measuring the sophistication of the reasoning behind these recommendations. To the extent that deliberations reveal participants' reasoning, the text of these deliberations could be analyzed for sophistication. Of course, some consideration may be needed for the possibility that not all participants will reveal their reasoning in such discussion. Alternatively, participants can be asked to provide details of their reasoning either in documents to be forwarded to the government body requesting the public input or in the context of pre- and post-discussion surveys. Regardless of the exact data capture strategy, such methods require researchers to analyze text for the sophistication of reasoning.

One approach to measuring sophistication is via hand coding content—having research assistants code content for indicators of sophistication in reasoning. While a few such content coding schemes have been proposed [9], little evidence exists that these schemes capture conceptual sophistication and none are widely recognized as definitive measures. DeER project experiences with direct content coding for quality of reasoning, based on our content coding scheme [20], clarifies the difficulty of this task. Despite extensive effort, a liberal and a conservative coder could not come to agreement on the quality of ideologically charged arguments. Apparently, the perceived quality of such argument depends heavily on the background understandings of the coder. When people discuss political issues, they do not present their arguments as would mathematicians: spelling out their premises and taking one logical step after another on their way to a conclusion. Points made are sketches dependent on a much larger body of hidden understandings, premises, and reasoning. A liberal coder viewing a liberal argument will typically understand the hidden components behind the argument and therefore rate it more highly than would a conservative coder.

Another difficulty is that the reasoning present in a person's discussion is often spread over many disconnected passages of speech. Hand coding generally depends on classification of small passages and therefore has difficulty assessing the quality of a given passage when that quality depends on the full body of what a respondent says.

Automated approaches to capturing sophistication have also been developed, but their applicability to public discussion of policy issues is unclear. One approach involves rating speech by the number of words a speaker uses that indicate uncertainty and flexibility in considering alternatives [5]. This approach, developed by Margaret Hermann for studying political leaders, holds that more sophisticated leaders are those who are more flexible and willing to embrace uncertainty. The approach may not be particularly applicable to public discussants who are coming to grips with complex issues, in which expressions of uncertainty may well not be indicators of sophistication.

1.2.Toward Better Hand-Coded Measures

One promising possibility with respect to hand-coded measures of deliberative quality involves focusing not on argument quality directly but on objectively measurable verbal behaviors that suggest a high degree of engagement. In this paper, we present measures of the number of times participants raised questions, disagreed with others, introduced a new idea to a conversation, and suggested another topic for discussion. Such verbal behaviors are moderately straightforward to measure with high intersubjective agreement and help capture intellectual activity on the part of participants.

We present evidence that counts of verbal engagement behavior capture deliberative quality using a strategy of verifying the convergent validity of this and alternative indicators. These alternative indicators including two measures of topic knowledge and an automatic measure of conceptual connection. These are measures of what might appear to be quite different things—but for their common connection of seeking to capture deliberative quality. In addition, the indicators utilize quite different measurement strategies—interpretive content analysis, statistical language analysis, and survey methods. Again, if there is a common correlation among these measures, it would seem likely only to the extent that they capture deliberative quality.

In addition to hand-coding indicators of intellectual engagement, we are also pursuing more direct measurement of reasoning quality via instances of reasoning types identified in part in argumentation theory [18]. For example, our participants utilized metaphors and analogies, hypotheticals, references to each other's arguments, and other types of reasoning. These reasoning types may be counted with adequate reliability and might themselves be rated for degree of sophistication. A combined score might then be constructed that reflects both the number of times a participant presents reasons and the quality of that reasoning—a weighted sum with weights adjusted to reflect power to predict other indicators of deliberative quality. While we are currently pursuing such a measure of argument quality, it was not far enough along to include in this paper.

1.3.Toward a Conceptual Connection Measure

One approach to the measurement of reasoning sophistication used widely in psychology, particularly to study political reasoning, is integrative complexity [21,22]. In this theoretical framework, reasoning is abetted by differentiation and integration of concepts. People are viewed as more sophisticated to the extent that they have a wide variety of different dimensions or perspectives to bring to bear on a topic but also to the extent that they have integrative concepts—that is, they possess more general concepts that suggest relationships among the differentiations. For instance, the concepts of "liberal" and "conservative" help to integrate a substantial portion of the differentiated dimensions of the political realm. The theory of integrative complexity has been fruitfully applied to political speeches and texts [22]. Questions remain, however, as to whether the ideological perspectives of coders will affect the reliability of this measure. And, of course, the method involves a considerable investment of time and effort.

We propose an automated method of measuring conceptual connections employing natural language processing and statistical techniques. This method measures sophistication in terms of the number and average density of connections between concepts in people's speech. A connection between concepts is operationalized as a statistically significant increase in the occurrence of one concept (C1) when another concept occurs in a unit of text (C2). This definition corresponds to the conditional probability of C1 given C2: p(C1 | C2). We consider p(C1 | C2) "significant" to the extent that C1 occurs significantly more often in the presence of C2 rather than other concepts. In constructing this measure, natural language processing techniques help identify concepts from words.

A count of significant conceptual connections in participants' speech should help capture quality of reasoning, and hence deliberative quality, in a manner related to integrative complexity. In integrative complexity coding, differentiation is captured as the number of separate dimensions or perspectives discerned in the topic at hand. For instance, the integrative complexity coding manual counts each causal variable asserted to apply to a given issue as a separate dimension. Each statement making an assertion of a causal relationship between some variable and the issue at hand involves asserting a connection between the variable and the issue. This will likely manifest as a co-occurrence between one or a tuple (i.e., combination) of words that identifies the issue and one or a tuple of words that identify the causal variable. Thus, if many discussants assert this causal relationship, the terms should show at least a conditional probabilistic relationship—the presence of one of the terms should make it more likely to see the other. In particular, to the extent that the conversation is about the general issue at hand and is only in small part about the causal variable, p(mention of the causal variable | mention of the issue) is more likely to be significant than p(mention of the issue | mention of the causal variable).

Examining significant conditional relationships between words has a number of advantages. Hand-coded content analysis improves on simple counts of content words in part by assuring that the words used are actually substantively connected to the topic being discussed, rather than part of some tangent, and are being used to reason about the subject, rather than simply being mentioned in passing. Analysis of significant conceptual connections in participants' speech should make up some of the difference in measurement accuracy between simple word counts and hand-coded content analysis. The fact that two terms co-occur significantly across all speakers suggests that the terms are likely not idiosyncratic and are topical, assuming participants generally stayed on topic. Moreover, co-occurrence suggests the speakers are drawing connections between the terms, and thus are reasoning with them.

Conceptual connectedness may also provide a proxy for the integrative complexity measure of integration. Any given concept can have multiple connections both in terms of concepts that become more probable in its presence or concepts that make it more probable. Concepts with more connections should be more integrative, thus corresponding to the notion of integration concepts in integrative complexity theory.

2.METHOD

2.1.Participants

Participants were 53 student volunteers from classes in the Dept. of Communication at the University at Albany, SUNY. Participants were 54% female; 70% Caucasian, 8% Hispanic, 7% Asian, 4% African-American, 5% other, and 6% refuse to answer. The median age was 22 with most participants 21 through 22. Students reported that their family's income was $60,000-$90,000. Participants are similar to the type of people who tend to participate in American politics: Caucasian, upper-middle class, and college educated. Because they are older than the average college student, participants may be more representative of the non-student population.

2.2.Procedures

Professors of Dept. of Communication classes agreed to announce the project in their classes and provide extra credit to those who completed participation. An alternative assignment with equal extra credit was made available. Interested students were asked to register on a project website. Once registered, they received emails that asked them to complete a pre-discussion survey before a specific date. The end of the survey provided a web address for information regarding the network neutrality regulation topic and for the discussion boards. Those who completed the survey also received reminder emails with this web address. An email announced the beginning of online discussion which took place over the course of two weeks. Once discussions were completed, participants were asked to complete a post-discussion survey and some were asked to participate in focus groups to evaluate the project.

2.3.Measures

2.3.1.Automated Conceptual Connectedness Measure

The methods in this paper are oriented to extracting meaning from text by identifying how people use words and, ultimately, concepts. Of course, many of the words people use are merely linguistic scaffolding and therefore do not correspond to concepts. Also, the same word can be used with different meaning. The word "account," for example, can signify a bank account, an account of an event, whether a person can account for an action, or a "Millennium Account." Notice that these uses of the word correspond to different parts of speech, respectively: noun, noun with a different meaning, verb, and proper noun. We can, consequently, disambiguate the meaning of a word to a degree by learning its parts of speech role. Information about each word's parts of speech also allowed us to remove the many words in the scaffolding of language that are unlikely to correspond to concepts. Parts of speech classifications provided by NLP software were used to narrow our list of words to nouns and adjectives (which qualify nouns) and a handful of words in other categories that seemed pertinent. Furthermore, we narrowed our effort to the 171 nouns and adjectives that occur at least 10 times in all the messages. Below this number, statistical analysis is unlikely to turn up relationships.

Identifying words' parts of speech can be an important step in the direction of identifying particular concepts in people's speech. In particular, the verb form of many words have quite different meaning than the noun form. Fortunately, computer scientists have developed highly accurate methods of identifying a word's part of speech role.

We utilized the open-source General Architecture for Text Engineering (GATE) to process the messages participants left on the discussion bulletin board. In addition to its capacity to tag words with their parts of speech, GATE [3] also performs a number of other useful functions, such as clarifying where words and sentences begin and end, identifying punctuation, and so forth. We were able to use the structured data that GATE returns to mark word tuples—that is, combinations of words such as "net neutrality."

Statistical bootstrapping [6] was used to determine which concepts have statistically significant conditional relationships. Bootstrapping makes no distributional assumptions but rather infers the distribution from resampling of the data itself. Freedom from distributional assumptions is key because the distributions of words may be very complex—affected by contingencies in the speaking context and the context of sentences, paragraphs, and messages. Indeed, it is unlikely that speakers draw from the same "urn of words" across messages.

With 171 words, the search for significant conditional relationships among words encounters over 29,000 different combinations. Thus, one concern is whether the relationships that appear to be significant may only be due to coincidental relationships uncovered by a very large number of tests. A Bonferroni correction, in which the p-value is multiplied by the number of tests would, however, be inaccurate and extremely conservative. This correction makes sense in the context of independent p-values. In the analysis of conditional relationships, a given word occurs significantly more often in the presence of another word only because it occurs less frequently in the presence of yet other words. Thus, the p-values are not independent. In fact, the vast preponderance of conditional relationships in this study proved significantly non-related: a word does not occur at all in the presence of another word or very rarely. This is partly a function of the fact that speech often consists of quite a diversity of words. Conversely, to the extent a word that never occurs with most other words does repeatedly occur with a given word suggests a real relationship. Thus, this study will consider all conditional relationships with a p-value at or below .01 as significant. A more accurate indicator of the multi-test p-value is possible but time consuming to implement and not central to the purpose of this paper, which will independently ascertain the value of an overall indicator of the number of conceptual connections by testing this indicator for significant relationships with other measures of sophistication.

The conceptual connectedness measure (also called "cumulative conceptual connectedness") is constructed by counting the number of instances in which a given person uses two terms that are significantly connected in the body of all messages. With a larger data pool it may be advantageous to determine whether terms are significantly related within individuals. Here, however, it is sufficient to stipulate that if terms are significantly related, their connection serve some real function in the discussion. Because most terms do not occur at all with most other terms, even one instance of a co-occurrence of terms that co-occur significantly overall is likely meaningful.

We focus on the co-occurrence of two terms in a message rather than, for example, in a sentence. This parallels the unit of analysis in integrative complexity coding, which examines units of text that express a complete thought, typically paragraphs. A brief bulletin board message typically does express one thought. A differentiation that associates, for instance, a causal variable with an issue of discussion might occur over more than one sentence. To the extent that significant associations occur at the sentence levels, these should typically be detectable at the message level. Of course, looking at the wider context of the message might increase the possibilities for coincidental co-occurrence of terms, but a sufficiently rigorous alpha-level for significance should make accepting coincidences improbable.

One further constraint on the conceptual connectedness measure is that it only includes terms that were *a priori* determined to be relevant to an intelligent discussion of policy. Some words involved in highly significant conditional relationships do not bear in any direct way on policy discussion—for example, "I" and "able." Terms such as "control," "free," "speech," "right," and "fee" are directly relevant to a discussion of network neutrality policy. It will be necessary to determine whether any significant relationships between the conceptual connectedness measure and other measures of sophistication might simply be the result of counting policy-relevant words.

We will not here examine a measure of the integrative quality of respondents' words, such as respondents' use of words involved in large numbers of conditional relationships. Discussants used very few integrative terms that bear on policy. This is likely a result of the unfamiliarity of participants with the issue at hand.

A final consideration is whether the conceptual connectedness measure should be used as is or should be adjusted by the total number of words a respondent wrote. The density of conceptual connections, even in small amounts of text, may capture sophistication. On the other hand, any given co-occurrence of significantly related words is only a weak sign of sophistication. Confidence that a person is indeed sophisticated rises when the person displays more connected speech over larger bodies of text. A small amount of text with high connectedness could indicate the speaker was parroting more sophisticated conversation, as may have happened when some of our participants, scrambling to acquire extra credit, first entered the discussion near its end. High variability in amount of text per person suggests that a count of conceptual connectedness across all text may be a better measure than an indicator of average sophistication if one indicator is chosen. It will be important to determine that validating relationships cannot simply be explained by total word count.

Cumulative conceptual connectedness and average conceptual connectedness may play a role together in determining sophistication. There could be trade-offs between these measures. For example, a person who speaks a large amount with low average connectedness may have a relatively high cumulative connectedness score, but may not be very sophisticated. People who speak little with high average connectedness may be parroting or summarizing others.

2.3.2.Hand Coding

Hand coding was conducted in line with a previously published content coding scheme and procedure [20]. The student deliberations were content analyzed following guidelines of content analysis established by Krippendorff and Neuendorff. A codebook was developed that described the categories to be coded and decision rules. Three coders were trained. Analysis of intercoder agreement was conducted using Krippendorff's Alpha, and most measures are at the minimum or exceed the generally accepted thresholds for satisfactory agreement. Coders found difficulty in achieving a satisfactory intercoder agreement for identifying the topic and the valence of the message, and so topic and valence were coded by all coders, then differences were discussed and reconciled.

Hand coding measures are used to show convergent validity with the automated measure. Four coding categories were deemed pertinent as indirect measures of sophistication: the speaker asked questions of other participants ("Question"), the speaker disagreed with other participants ("Disagree"), the speaker introduced an argument that offers another way to address the problems that are meant to be addressed by the proposed network neutrality regulation new ideas ("New Idea"), and the speaker introduced a topic unrelated to those focused on network neutrality (suggesting they were off topic). ("Other Topic"). Each instance of these speech acts was added up for each participant to create a score for that participant.

2.3.3.Survey Indicators

Three survey scales were also included in these analyses. One scale, Neutrality Knowledge, is a score from a multiple-choice quiz in which respondents were asked to correctly answer factual questions about the discussion topic. For instance: "Net neutrality is...." or "What are 'common carriers'?" A second scale, "Self-Reported Knowledge," asked participants to rate their own level of knowledge with regard to the topic. Self-reports might capture aspects of knowledge not captured by the multiple-choice quiz. Finally, the "Linear Government" scale indicates the degree to which the respondent agrees with a series of statements indicative of understanding the government in terms of a simple, linear command hierarchy rather than as a complex system of checks and balances. Politically more sophisticated respondents should have low scores on this measure.

2.3.4.Control Indicators

The total word count for a participant (Total Words) is used as a control variable for certain analyses. In addition, to determine whether conceptual connectedness was significant only because it counts policy-relevant terms, another more complete indicator of policy terms was constructed (Policy Terms). While conceptual connectedness only counts policy-relevant terms that have significant conditional relationships with other policy-relevant terms, Policy Terms is a count of all policy-relevant words in the full set of 171 words.

3.RESULTS

3.1.Convergent Validity—Conceptual Connectedness

Table 1 shows the correlation of conceptual connectedness with several other indicators of sophistication and the significance of these correlations. All correlations are in expected directions, hence p-values are taken as one-sided. Conceptual connectedness proves to have significant relationships with four of the seven variables, including the objective measure of neutrality knowledge and self-reported knowledge. It also correlates significantly with the human content coding measures such as the number of questions a participant asks of others and how often the participant disagrees with others. Connectedness shows trend relationships with proposing new topics and linear conceptions of government. While not depicted, average connectedness correlates significantly with the latter (ρ=-.24; p=.045).

Table 1. Correlation of Conceptual Connectedness with Other Indicators of Sophistication and Their Significance

Other Indicators	**Conceptual Connectedness** **ρ (p-value, one-sided)**
Neutrality Knowledge	.33 (.008)
Self-reported Knowledge	.35 (.006)
Question	.45 (.0003)
Disagree	.38 (.002)
New Idea	.13 (.17)
Other Topic	.20 (.07)
Linear Government	-.19 (.09)

Notes: N=about 53 throughout.

3.2.Alternative Explanations

Perhaps conceptual connectedness significantly relates to other sophistication measures only because it indirectly captures how much a participant wrote. Those who write more may be more informed. Separating these two explanations, however, may not be straightforward. Conceptual connectedness is the product of the total number of words and the average connectedness per word. The contribution of conceptual connectedness beyond total word count is average connectedness, which, for reasons described above, may not constitute a particularly good measure of sophistication. The one other contribution connectedness may make above and beyond total word count are trade-offs between cumulative conceptual connectedness and average connectedness. These, then, become relevant in pitting connectedness against total words as explanations of sophistication.

Table 2 examines regressions of neutrality knowledge, both objective and self-reported, on total words, connectedness, and average connectedness. The regressions are that of interaction effects: total words + average connectedness + average connectedness X total words. The latter term is cumulative connectedness. This interaction form allows for nonlinear trade-offs to emerge between average and cumulative connectedness.

The first column of Table 2 shows that none of the substantive independent variables prove statistically significant. The F-test of the regression, however, is significant, suggesting that the variables are too collinear or the sample size is too small to determine which significantly affects the outcome. The two connectedness variables show lower p-values than total words—an indication that total words does not have the upper hand. For self-reported knowledge, both connectedness variables are significant, while total words proves non-significant. Self-reported knowledge is more than simply a subjective delusion, given that it significantly correlates with the objective knowledge measure (p=.01).

The coefficients for the regression of self-reported knowledge show that this indicator is greatest in the case that both total words and average connectedness are relatively high. Keep in mind that cumulative connectedness is average connectedness times total words. The meaning of the regression can best be understood by plugging various values into the regression equation. Inserting first quartile values of total words and average connectedness, the total effect of total words and average and cumulative connectedness is -3.76. Self-reported knowledge was measured on a seven-point scale, so this proves to be a very substantial reduction. In contrast, with both variables set at third quartile levels, the total effect on self-reported knowledge is +2.7, a substantial increase.

Table 2. Ordinary Least Squares Regressions Showing Effects of Total Words and Connectedness on Neutrality Knowledge, Objective and Self-Reported

Independent Variables	**Dependent Var.** **Neutrality Knowledge** **Unstd. Coef (two-sided p-value)**	**Dependent Var.** **Self-Reported Knowledge** **Unstd. Coef (two-sided p-value)**
Total Words	.06 (.75)	-.002 (.17)
Average Connectedness	-1.14 (.33)	-18.5 (.02)
Cumulative Connectedness	.002 (.59)	.05 (.03)
Constant	.59 (.0001)	2.3 (.0002)
R^2; s.e.; N	.15; .21; 52	.22; 1.39; 52

One question is whether an indicator of the total number of policy words used by a participant would do better than connectedness. Policy words was analyzed in a manner similar to Table 2, because it is possible to consider average number of policy words or cumulative number. Unlike with connectedness, none of the coefficients proves significant.

3.3.Convergent Validity—Hand-Coded Content Analysis

As with the conceptual connectedness measure, the hand-coded measures of verbal engagement behaviors prove to have a variety of intercorrelations. The Question, Disagree, New Idea, and Other Topic variables are well-correlated among themselves, with a minimum correlation of .42 (p=.001). These intercorrelations among separate behaviors suggest they measure something in common, hopefully deliberative quality. Table 1 shows that three of these measures are significantly related to the quite different conceptual connectedness measure. Finally, Table 3 finds that the Question measure correlated significantly with objectively-measured net neutrality knowledge. Also three of the variables significantly correlate with self-reported knowledge and the fourth shows a trend. None of the hand-coded variables significantly correlate with linear notions of government.

Table 3. Correlation of Hand-Coded Measures with Other Indicators of Sophistication and Their Significance

Hand-Coded Measures	Neutrality Knowledge **ρ (p-value, one-sided)**	Self-reported Knowledge **ρ (p-value, one-sided)**
Question	.37 (.004)	.25 (.05)
Disagree	.14 (.15)	.19 (.09)
New Idea	.15 (.14)	.30 (.02)
Other Topic	.10 (.23)	.42 (.001)

Notes: N=about 53 throughout.

4.DISCUSSION AND CONCLUSION

Our hand-coded measures of verbal engagement behaviors demonstrated good convergent validity, which suggests they may be useful for capturing deliberative quality. These measures correlate well among themselves, which implies they may measure something in common, such as deliberative quality. They also correlate with very different types of measures of deliberative quality, including conceptual connectedness, objectively-measured factual knowledge of network neutrality, and especially self-reported knowledge of net neutrality. Self-reported knowledge significantly correlates with an objective knowledge measure and may capture aspects of knowledge not reflected in the objective measure, which was fact-focused.

We find that the automatically coded conceptual connectedness measure examined here possesses considerable convergent validity with a variety of indirect indicators of sophistication of reasoning. These include topical knowledge, hand-coded content measures of participants' pro-activeness during discussion, and sophistication of views of government. The measure is, however, collinear with the total number of words participants used, which might also capture sophistication. Not surprisingly, given the small dataset and strong collinearity, conceptual connectedness did not clearly differentiate its value as distinct from total word count, except in the case of self-reported network neutrality knowledge.

This evidence on behalf of the connectedness measure suggests it will be worthwhile to continue to pursue this approach to measuring sophistication. A larger dataset could prove helpful by allowing analysis of the significance of conditional relationships by individual rather than assuming, as we did here, that relationships that are significant across all persons hold for particular individuals. A dataset with more sophisticated discussants should make it fruitful to develop a related measure of the integrative properties of words used. Rather than simply counting the number of times a person uses pairs of words that are significantly conditionally dependent, a measure of integration would weigh these counts by the number of connections each word has with others. Such a measure should be less strongly related to total word counts than the measure considered here. Subsequent research in the DeER project will pursue developing such datasets.

While not the focus of this paper, the hand-coded measures of sophistication appear very promising. They significantly correlate with neutrality knowledge while having moderate correlation with total word count.

5.ACKNOWLEDGMENTS

This research is based upon work supported by the National Science Foundation under Grant No. IIS-0713143. We would also like to thank Nick Webb, whose efforts made this project possible.

6.REFERENCES

1. Benkler, Y. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, 2007.

2. Button, M. and Ryfe, D.M. What Can We Learn from the Practice of Deliberative Democracy? In J. Gastil and P. Levine, eds., *The Deliberative Democracy Handbook: Strategies for Effective Civic Engagement in the Twenty-First Century*. Jossey-Bass, San Francisco, 2005, 20-34.

3. Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. Gate: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Annual Meeting of the ACL*, (2002).

4. Dahlberg, L. The Internet and Democratic Discourse: Exploring The Prospects of Online Deliberative Forums Extending the Public Sphere. *Information, Communication and Society 4*, 4 (2001), 615-633.

5. Dyson, S.B. Text Annotation and the Cognitive Architecture of Political Leaders: British Prime Ministers from 1945-2008. *Journal of Information Technology & Politics 5*, 1, 7-18.

6. Efron, B. and Tibshirani, R.J. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

7. Fishkin, J.S. *Democracy and Deliberation: New Directions for Democratic Reform*. Yale University Press, New Haven, 1991.

8. Gastil, J. *Political Communication and Deliberation*. Sage Publications, Inc, 2008.

9. Graham, T. and Witschge, T. In Search of Online Deliberation: Towards a New Method for Examining the Quality of Online Discussions. *Communications (Sankt Augustin) 28*, 2 (2003), 173-204.

10. Habermas, J. *The Theory of Communicative Action, Volume Two: Lifeworld and System: A Critique of Functionalist Reason*. Beacon Press, Boston, 1984.

11. Kriplean, T. Designing Mediating Spaces Between Citizens and Government. .

12. Luskin, R.C., Fishkin, J.S., and Jowell, R. Considered Opinions: Deliberative Polling in Britain. *British Journal of Political Science 32*, 3 (2002), 455-488.

13. Muhlberger, P. Human Agency and the Revitalization of the Public Sphere. *Political Communication 22*, 2 (2005), 163-178.

14. Muhlberger, P. and Weber, L.M. Lessons from the Virtual Agora Project: The Effects of Agency, Identity, Information, and Deliberation on Political Knowledge. *Journal of Public Deliberation (available at http://services.bepress.com/jpd/) 2*, 1 (2006), 1-39.

15. O'Connor, R.E., Bord, R.J., Yarnal, B., and Wiefek, N. Who Wants to Reduce Greenhouse Gas Emissions? *Social Science Quarterly 83*, 1 (2002), 1-17.

16. Petty, R.E., Wegener, D.T., and Fabrigar, L.R. Attitudes and Attitude Change. *Annual Review of Psychology 48*, (1997), 609-647.

17. Rheingold, H. *The Virtual Community: Homesteading on the Electronic Frontier*. Addison-Wesley Pub. Co., Reading, MA, 1993.

18. Rottenberg, A.T. *The Structure of Argument*. Bedford/St. Martin's, 2002.

19. Sears, D.O. and Citrin, J. *Tax Revolt: Something for Nothing in California*. Harvard University Press, Cambridge, Massachusetts, 1982.

20. Stromer-Galley, J. Measuring Deliberation's Content: A Coding Scheme. *Journal of Public Deliberation 3*, 1 (2007).

21. Suedfeld, P., Tetlock, P.E., and Streufert, S. Conceptual / Integrative Complexity. In C.P. Smith, J.W. Atkinson and E. al, eds., *Motivation and Personality: Handbook of Thematic Content Analysis*. Cambridge University Press, New York, NY, US, 1992.

22. Tetlock, P.E. Cognitive Structural Analysis of Political Rhetoric: Methodological and Theoretical Issues. In S. Iyengar, ed., *Explorations in Political Psychology*. Duke University Press, Durham, NC, 1993, 380-405.