Political Psychology

Political Psychology, Vol. 35, No. 5, 2014
doi: 10.1111/pops.12207

Integrative Complexity Coding Raises Integratively Complex Issues

Philip E. Tetlock
University of Pennsylvania

S. Emlen Metz
University of Pennsylvania

Sydney E. Scott
University of Pennsylvania

Peter Suedfeld
University of Pennsylvania

Conway, Conway, Gornick, and Houck (2014) report a major effort to automate integrative complexity coding. Judging this effort requires researchers to be more explicit in articulating key methodological assumptions about the coding process and theoretical assumptions about the construct. Unresolved issues include: (1) when, and on what basis, we should attribute divergences between human coders and algorithms to overestimations or underestimations by one or the other approach; and (2) to what extent second-generation algorithms can yield Pareto improvements that reduce errors of both underestimation and overestimation. Further progress in developing natural language processing measures of this cognitive style will require sharper definitions of target constructs: in particular, different types of differentiation (dialectical and elaborative) and integration (hierarchical and flexible) and clearer guidelines for factoring context into assessments.

KEY WORDS: integrative complexity, cognitive style

Introduction

Assessing even simple political-psychological constructs from natural language is a nontrivial undertaking, whether the codes are assigned by trained human coders or automated computer programs. Assessing complex psychological constructs, such as integrative complexity, is even more daunting, particularly for automation. For example, human coders can assess evaluative content in a text and notice contrasting values applied to the same object with relative ease. There is little doubt that the speaker disapproves of "running dog, capitalist lackeys," or "bullying, central planning bureaucrats." But it is far more difficult to create valid word count indicators of whether the speaker is endorsing a multidimensional or differentiated assessment of an attitude object, as occurs in integratively complex thinking. Using common qualifiers like "yet" or "however" often signals a new

Published by Wiley Periodicals, Inc., 350 Main Street, Malden, MA 02148, USA, 9600 Garsington Road, Oxford, OX4 2DQ, and PO Box 378 Carlton South, 3053 Victoria, Australia

dimension of evaluation ("a patriot, *however*, corruptible"), but this can be done without any such linguistic markers ("a patriot who proved corruptible"). Linguistic markers can even signal the amplification, not the qualification, of initial evaluation ("a bully, however you look at it"). Word count programs therefore must rely upon the probability that each word reflects increased integrative complexity of a text, which may vary across contexts and renders codes necessarily coarse.

These complexities have led most integrative complexity researchers to rely on well-trained human coders to make the necessary natural language distinctions (e.g., Feist, 1994; Gruenfeld, 1995; Gruenfeld & Hollingshead, 1993; Levi & Tetlock, 1980; Pancer, Hunsberger, Pratt, Boisvert, & Roth, 1992; Porter & Suedfeld, 1981; Suedfeld, 1985; Suedfeld & Bluck, 1993; Suedfeld & Rank, 1976; Suedfeld & Tetlock, 1977; Tetlock, 1979, 1981, 1984: Tetlock & Tyler, 1996; Winter, 2007). The methodological-epistemological trade-offs between reliance on human judgment versus automated algorithms are fairly obvious. Trained human coders can make more subtle distinctions that algorithms might miss. But human coders can be expensive, time-consuming, and sometimes erratic (lowering reliability coefficients) or systematically biased by, for instance, playing favorites and assigning higher scores to speakers whom they like. Computer algorithms offer the promise of making integrative complexity coding conform to neo-positivist norms of science; the algorithms are perfectly transparent and reproducible. But do they capture the key construct that we set out to capture? Are we falling prey to the drunkard's search—looking for our keys under the lamppost because that is where it is easiest to see?

Of course, we never know what is possible until we try—and researchers such as James Pennebaker (Tausczik & Pennebaker, 2010), Margaret Hermann (1999) and Michael Young (2001) are pioneers in developing automated measures of complex psychological constructs from natural language data. This symposium focuses on the most recent effort to capture a complex construct—integrative complexity—via computer algorithms (Conway, Conway, Gornick, & Houck, 2014).

The structure of the symposium is simple. Conway et al. have ably sketched what has been accomplished. Building on this, we offer our views of lingering conceptual puzzles that, if not systematically addressed, will limit the long-term potential of the research program.

The Allure of Automation

As integrative complexity researchers well know, training coders and ensuring they remain reliable throughout the text analysis process is a labor-intensive undertaking. We agree with Conway et al. (2014) that the labor costs of human coding of integrative complexity have probably contributed to its underutilization in recent years. Automation promises to eliminate this drain on resources as well as to short-circuit endless debates among coders on questions that can tax even Talmudic temperaments: How implicit must differentiation be to qualify as a "2"? And how explicit must it be to qualify as a "3"? How implicit must integration be to qualify as a "4"? And how explicit must it be to qualify as a "5"? Thus, one of the most alluring appeals of automation is that algorithms can be efficiently applied to vast volumes of text, allowing for it to be more widely used and potentially compensating for some loss of nuance in particular cases.

Automation would also bestow a new aura of scientific status on the construct; it represents a crucial step from our current intersubjective measurement standard (agreement among trained coders) to an objective measurement standard. An automated system would provide perfectly specifiable and replicable measurement operations that do not require initiation into a specialized epistemic community. Although researchers would need to keep an eye on changing habits of speech and the consequently inevitable slippage between any automated system and the forms of language conveying integrative complexity (a fundamentally hermeneutic construct), an automated system provides greater explicitness and regularity in measurement.

Nonetheless, if it were easy to automate integrative complexity coding, someone would have already done it. Numerous teams over the last 50 years have been conducting integrative complexity research, and, as these teams know, there are good reasons for worrying about that: (1) Algorithms will often underestimate or fail to accurately recognize integrative complexity; and (2) algorithms will often overestimate integrative complexity, lured by specious signs of differentiation and integration that skilled human coders would have seen through.

Identifying Errors and the Elusive Quest for a Gold Standard

If human judgments of integrative complexity and algorithm classifications were perfectly aligned, the transition from a more subjective to more objective measurement standard would be seamless. But, as Conway et al. (2014) show, even the best correlations between human IC codes and automated IC codes are only moderate—and even these values almost certainly understate the divergence (e.g., large correlations between two sets of codes can comfortably coexist with poor agreement between sets, as pointed out by Young & Hermann [2014] and with large mean differences between scores from one set and from another set).

Of course, divergences need not always be the fault of the algorithms. Divergences could arise because either the humans or the algorithms overestimate or underestimate complexity in coding either differentiation or integration—errors that the other approach avoids. It is instructive to explore examples of each—instructive because we quickly discover the need to be more explicit about the very definition of integrative complexity. What counts as an error hinges on exactly what we mean by integrative complexity.

Distinctive Vulnerabilities of Algorithms

Overestimation algorithm errors in algorithms. The algorithm increases differentiation scores whenever the speaker uses key words such as "however" or "yet." But speakers do not always use these linguistic markers to introduce a new dimension of evaluation of the attitude object. Sarcastic and ironic passages often use these linguistic markers of differentiation in a specious manner. Furthermore, most words which often indicate differentiation are also commonly used in ways that do not. For example, the highly differentiating word "however" sometimes not only fails to indicate differentiation, but, on the contrary, indicates a stronger unidimensional claim, as in Thomas Mowbray's claim of loyalty in Shakespeare's *Richard II*: "*However* heaven or fortune cast my lot/There lives or dies, true to King Richard's throne, a loyal, just, and upright gentleman" (I.iii.365–367). Conway's automated IC algorithm codes this statement as a 2.5, indicating semiexplicit differentiation where there is none at all.

The algorithm will also inflate integration scores when the speaker uses polysemous keywords such as "integration" or "interaction" but has no intention of using these words to signal the synthesis of competing considerations. For instance, the speaker might say, "The European monetary *integration* was a stupid idea 20 years ago, and it is even dumber now," and the automated IC algorithm would assign a score of 3.5. This is halfway between explicit differentiation and implicit integration; human eyes, however, tell us neither is present.

Underestimation errors of algorithms. The algorithm may fail to recognize differentiation because the speakers do not always use any particular words to signal a new dimension of evaluation. Existing algorithms, for example, are hard-pressed to detect the glaringly obvious differentiation in the opening lines of Charles Dickens's *A Tale of Two Cities* (which is assigned a measly score of 1.5 by the automated IC algorithm): "It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter

of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way . . .” (Dickens, 1859/2003, p. 1).

Likewise, the algorithm may fail to recognize integration when a speaker does not use keywords for integration, even though the speaker is synthesizing competing considerations. Such a synthesis could take a succinct mathematical form, e.g., “$e = mc^2$” (assigned a 1 by the automated IC algorithm, indicating no differentiation at all), or a more discursive form, e.g., “Game theory tells us that quasi-randomly shifting from backhand to forehand to backhand is a sound strategy in tennis” (also assigned a 1 by the automated IC algorithm).

Distinctive Vulnerabilities of Human Coders

Overestimation errors of human coders. Insofar as human coders are predisposed to be generous in assigning higher complexity scores to speakers whom they like or causes they endorse (Conway et al., 2014), they will often see signs of differentiation in the absence of the linguistic markers for algorithms. For instance, they might assign a high differentiation score to a favored candidate engaging in what Jervis has aptly called belief system overkill: “during my years in office, unemployment has declined, inflation has been low, and productivity has grown.” The manual for human integrative complexity scorers warns against assigning “differentiation” scores to a mere list, but biased scorers might overlook this point. Furthermore, although the preparation of materials for scoring tries to prevent the identification of the source, this is not always possible. However, the automated IC algorithm correctly codes this passage as a 1, indicating no differentiation.

Just as sympathy for the speaker can inflate estimates of differentiation, it can also inappropriately inflate estimates of integration: “We believe that only capturing many points of view at the highest levels will solve the problems confronting this country.” The automated IC algorithm, unswayed by emotional resonance of the content, assigns this statement a 1.

Underestimation error in human coders. The flipside bias among human coders is a tendency to be too stingy in assigning higher complexity scores to speakers whom coders see as unsympathetic. Imagine a high-level Nazi official discussing the logistical challenges of managing the Holocaust during World War II: “We want to kill as many Jews as quickly as possible; however, we also don’t want to make exorbitant demands of transportation infrastructure also needed for winning the war.” The automated IC algorithm assigns this passage a 2.5, capturing the tension between two goals that morally outraged human coders might well miss.

Again, just as animus toward the speaker can deflate estimates of differentiation, it can inappropriately deflate estimates of integration. Imagine that the Nazi official has continued: “The right balancing requires close coordinating between the SS and Wehrmacht high command to identify the least disruptive-to-the-war-effort tracks and times for transporting inferior races to death camps.” The officer has integrated by maximizing one function subject to constraints specified by the other function, and the automated IC algorithm assigns a score of 3.5. Even aside from the problem of bias, human scorers often underestimate high levels of complexity, playing it “safe” because it can be difficult to judge whether a passage meets the criteria for a score above 5. This tendency may contribute to the small number of 6 and 7 scores found in most sets of data.

The Quest for a Gold Standard

From a positivist perspective, the most disconcerting feature of these examples is the absence of any “gold standard” that can serve as the ultimate arbiter for the human or algorithmic coding of integrative complexity. What should be the ultimate resolution criteria for determining the “true integrative complexity” of a specific utterance in a specific context? Peter Suedfeld has said inte-

grative complexity has elements of both an art and science—and we can now see clearly how far we are from the neo-positivist ideal of universal scientific metrics that carry the same meaning for everyone everywhere. In this spirit, in 1983, an international commission of metrologists officially defined a meter as the distance that light travels in a precise, extremely tiny, span of time, with no judgment calls or wiggle room (Crease, 2011). But it was not always so: there have been battles over standards stretching back over 200 years (the 1791 definition of the meter was one ten-millionth of the distance from the Equator to the North Pole at sea level). The scoring of cognitive complexity has not yet achieved a single consensual definition; its status is closer to the early years of the metric system. At this point in history, judgment calls dependent on context are unavoidable in assessing differentiation and integration.

The push to automate puts much greater pressure on integrative complexity theorists to refine conceptual distinctions in the very meaning of the construct. Automation highlights the necessity of distinguishing between dialectical differentiation, in which there is a genuine tension or ambivalence between perspectives, and elaborative differentiation, in which the speaker may be simply listing reasons why he or she is right and opponents are wrong (Conway et al., 2008; Tetlock & Tyler, 1996). It also highlights the value of distinguishing between hierarchical integration, in which the speaker offers a fixed interaction or trade-off rule for combining two perspectives, and flexible integration, in which the speaker recognizes the need to improvise different combinatorial rules in different situations.

More generally, if we are to consider new techniques for measuring integrative complexity, it is important to specify as precisely as possible what it is that we are trying to measure and accordingly our best criterion for evaluating the accuracy of a new measure. There are several possible approaches to establishing a gold standard criterion. (1) Duplicating the nomological net is a classic approach to validating a new measure of an established construct (Cronbach & Meehl, 1955). The nomological net describes the set of positive, negative, and nonsignificant relationships between the construct in question and other constructs/measures; if the IC algorithm fails to predict any established correlates of integrative complexity, it will not be of much use. (2) A more direct approach is assigning the role of ultimate arbiter of the score to some particular individual or group, such as (2a) the author, (2b) the intended audience, or (2c) the coder. We argue that a combination of (1) intelligent comparison of nomological nets between new and old measures and (2c) comparison to the scores of expert human coders will serve as the best possible criteria for any new measure of integrative complexity.

(1) Automated measures of integrative complexity lend themselves particularly well to comparison with the nomological net established in the last 40 years of human-coded integrative complexity, since applying the new computer algorithms is so time efficient. Several correlates of integrative complexity can be retested with large, publicly available sets of texts. For example, a number of studies in several labs have found higher integrative complexity among political leaders before engaging in successful negotiations and lower integrative complexity among political leaders before engaging in aggressive or risky action (Satterfield, 1998; Suedfeld & Rank, 1976; Suedfeld & Tetlock, 2001; Tetlock & Boettger, 1989; Tetlock & Tyler, 1996; see Conway, Suedfeld, & Tetlock, 2001, for a review; see Levi & Tetlock, 1980 for a possible exception). Large quantities of such political texts are publicly available, enabling researchers to test the coarser automated integrative complexity across a large dataset.

It would also be relatively simple to test the controversial but widely demonstrated correlation between integrative complexity and political orientation. Left, moderate, or majority-party politicians tend to exhibit higher integrative complexity than right, more extremist, or minority-party politicians (Suedfeld, Bluck, Ballard, & Baker-Brown, 1990; Tetlock, Hannum, & Micheletti, 1984; but see Van Hiel & Mervielde, 2003). Large swathes of text from each side of the political spectrum can be easily obtained and coded by the automated program. However, if automated algorithms do

not find an association, it is worth considering to what extent the established correlation might have arisen from bias among largely liberal human coders. It is thus possible that the automated system could illuminate biases in past research.

(2) At first blush, it makes sense to use the author or intended audience of a text as an ultimate (ideal) arbiter of the text's integrative complexity. Integrative complexity is a description of the structure of meaning in a text, and who knows better what is meant than the author? Or, if we admit that authors do not always succeed in saying what they mean, then surely the intended audience has the most perfect understanding of what the text conveys; moreover, there is evidence that people not versed in the theory or research nevertheless have an intuitive understanding of the integrative complexity construct (Suedfeld, de Vries, Bluck, Wallbaum, & Schmidt, 1996).

Despite these advantages, using either the author of a text or its intended audience shrinks our ability to use integrative complexity to compare texts *across* authors or audiences. The scoring and perception of integrative complexity are not value neutral. Authors and sympathetic audiences usually will wish to exaggerate their complexity; hostile audiences, to underplay it. Varying degrees of honesty and varying valuations of integrative complexity would lead to varying degrees of bias. Different construals of differentiation and integration would be inevitable, even across subcultures. Since comparability is the whole purpose of establishing a clear, consistent measure, these are fatal flaws.

What of the coders, the judges on whom we rely in practice? This is a logistical necessity, but fortunately it also makes good sense. Like authors and audiences, third-person coders may be biased by the content or context of a text, and unlike them, they lack privileged access to the intension of a text. However, the greater distance of third-person coders also leads to less bias. Even visual-perception scientists sometimes consider the perspective of an "ideal observer." We may train our coders to be as close to an ideal observer as possible; that is, one who is unbiased and culturally and historically informed about the broader context of the text, including references, conversational norms, and local differences in terminology. We can also minimize precise identification of the source and situation. Most importantly, trained coders can agree explicitly upon what kinds of context they should and should not take into account in coding a text. This is necessary if we wish to compare texts across contexts.

For this reason, some kinds of context should *not* be taken into account during the coding process, though they may be worth mentioning when comparing the texts. For example, considerable research has been done on the differences between representational and instrumental texts (a distinction very roughly parallel to intrapsychic vs. impression management). Several studies have compared the integrative complexity of public statements to private letters (Tetlock & Tyler, 1996) or of texts written under conditions of accountability to various judges versus anonymity (Tetlock, 1983). Higher instrumentality of a text (which is more likely to appear in deliberate public documents) could lead to either lower integrative complexity out of a desire to persuade the audience or to higher integrative complexity out of a desire to avoid offending an already opinionated audience. The former seems to have been true of Churchill's public statements about Nazi Germany in comparison to his private letters (Tetlock & Tyler, 1996). The latter seems to have been true of Japanese officials deciding whether to attack the United States, who were more integratively complex in discussions with the anxious Emperor than in their meetings with each other (Levi & Tetlock, 1980). The contextual differences between these texts were precisely what these studies wished to examine. Had coders tried to account for representational versus instrumental context in their codes, these provocative contrasts might never have been found.

It should be noted, however, that some studies have not found differences between private and public communications, and in some studies, complexity changes seem to be in the direction opposite to that which might be expected if complexity reflected a conscious decision to appear either conciliatory or uncompromising. For example, it would seem that a show of increasing

complexity would have been a sensible deceptive-impression management maneuver for leaders planning a surprise attack on another country, for Saddam Hussein facing impending sanctions from the UN Security Council, and for Mikhail Gorbachev as his economic policies were failing and his domestic political enemies gaining strength; yet in all of these cases, the complexity of the individual went down even while the overt content of many of the messages professed a fervent dedication to a negotiated peaceful resolution (e.g., Suedfeld & Bluck, 1988; Suedfeld & Rank, 1976; Suedfeld, Tetlock, & Ramirez, 1977; Wallace, Suedfeld, & Thachuk, 1993). One plausible inference is that changes in integrative complexity in these cases were valid indicators of the actual structure of thought rather than the results of deliberate deceit.

The Future of Algorithms

Current first-generation, word count algorithms imprecisely capture the integrative complexity of texts. Yet, meaningful increases in accuracy and precision are within reach of ambitious second-generation algorithms. In particular, natural language processing techniques hold considerable promise for developing an algorithm which more closely approximates human codes.

The Fuzziness of Complexity

It is useful to reconceptualize the task of coding integrative complexity in terms of fuzzy set theory. Whereas in classic or crisp set theory, membership to a category is all or none, in fuzzy sets elements have a degree of membership that can be zero (no membership), one (full membership), or any number in between. Fuzzy set theory is a particularly successful framework for natural language problems and human decision making, because natural language often has an "intrinsic fuzziness" characterized by imprecise or context-dependent meanings of a word (Zimmermann, 1991). "Creditworthiness," for example, is intrinsically fuzzy because it is often a matter of degree and moreover depends on multiple factors (financial bases, personality, judge's perspective).

Categories of complexity—"differentiated," "integrated," and "higher-order integrated"—are similarly fuzzy. First, the actual structure of thought in a passage can be differentiated or integrated to greater or lesser degrees, such that conceptualizing the set of "differentiated" statements and "integrated" statements as inherently fuzzy could ultimately allow more precise and theoretically appropriate measurement. Second, beyond the in-principle fuzziness of these categories of structures of thought, the communication of that structure of thought in natural language is so inherently imprecise that fuzzy set theory could provide a practically useful technique for capturing the imprecise information in codes. Human integrative complexity coding already captures some fuzziness. When passages "fall in the fuzzy boundary zone between scale values," they are assigned intermediate scores of 2, 4, and 6 (Baker-Brown et al., 1992, p. 402). In theory, viewing categories of complexity as fuzzy will eventually allow for a more precise and theoretically appropriate measure of complexity. Conway et al. (2014) use even finer-grained categories by including fractional integrative complexity scores such as 2.25, 2.5, 2.75, and so on.

Context broadening. One of the most frequent causes of divergence between human coders and first-generation algorithms is that the latter are "atomistic," focused on isolated words, whereas human coders consider textual and external context before assigning a score. Human coders may agree with the algorithm that, on average, "however" increases the likelihood of differentiation, but they are better positioned to spot when "however" is a specious semantic marker. For example, when Michelle Obama says, "My husband is a great president; however, he is also a great family man," human coders see from the context of the sentence that there is no ambivalence or dialectical differentiation indicated by "however." They can reasonably infer that "however" is a misleading indicator of (evaluative) differentiation in this context. Human coders are able to consider not only

the textual context but also the larger communication context. They can easily recognize that when a President's wife is praising her husband at a nomination convention, she is unlikely to express marital ambivalence when she uses "yet" or "however."

Our colleagues Young and Hermann (2014) have suggested that a multipass system can help ameliorate these difficulties, for instance, by coding different parts of speech differently. We also see potential in multipass systems, potential to incorporate more fine-tuned distinctions between word senses and indications of sentiments.

Progress in content analysis and automated systems offers possible solutions for incorporating context in automated coding. For example, latent semantic analysis and related techniques inductively infer word sense by using counts of words that co-occur with the target word (Hofmann, 2001; Landauer, Foltz, & Laham, 1998). Such techniques deal successfully with some problems that plague computers but not humans in coding integrative complexity. They can use training corpora to discriminate meanings of polysemous words (Hofmann, 2001; Schütze, 1998) and accurately match synonymous words (Landauer & Dumais, 1997).

Sentiment analysis is a still more promising technique. Sentiment analysis is a multipass algorithm which detects the sentiments of parts of a text (positive, negative, both, or neutral). As such, it could be used to detect differentiation of sentiments applied to a single object, which first-generation systems are likely to miss (Wilson, Wiebe, & Hoffmann, 2009). Although there are many versions of this technique, a particularly promising one estimates the sentiment of a phrase based on its estimated "prior" sentiment in a separate lexicon and then adjusts sentiment based on context of the other nearby words in the clause or sentence (Wilson, Wiebe, & Hoffmann, 2009).

Consider our previous example: "My husband is a great president; however, he is also a great family man." Now consider the variant: "My husband is a great president; however, he is also a great liar." This variant, unlike the original, exhibits evaluative differentiation. It would be easy to judge the difference as a human coder, but Conway's system assigns the same score of 2.5 to both sentences. Sentiment analysis, however, would be able to tag the phrase "a great liar" as connoting negative sentiment and "a great president" a positive sentiment. The sentence could thus be classified as "differentiated" based on the conflicting sentiments within the sentence. Other techniques may enable an algorithm to recognize when sentiments are being applied to the same or to different attitude objects. It may even be possible to use other sophisticated machine-learning techniques to aggregate a variety of natural language processing algorithms for the most accurate integrative complexity codes. We can use Conway et al.'s work as a foundation upon which to collaboratively build a more sophisticated integrative complexity coding algorithm, utilizing the work of other natural language processing experts.

This analysis obviously just skims the surface of the underlying complexities of integrative complexity coding—and the massive conceptual challenges confronting future efforts to create successful automated coding systems. It also underscores the great value of such endeavors.

We have spilled so much ink discussing the ideal human coder in an article about automation because, however transparent, objective, and replicable an automated coding system is, it must ultimately be measured against the best standard of human coding. Automated systems can only measure behavior, i.e., the patterns and structures of words and phrases. Integrative complexity refers to the structure of *meanings* in a text, of which the structure of words can never be more than a context-dependent proxy. Even if we should reach a point where the algorithmic model of human judgment exceeds human judgment in accuracy—as indexed, say, by larger validity coefficients for well-replicated effects—the algorithm should still be checked against human judgment periodically to detect slippage between the algorithmic measure and the meaning expressed, as language changes across domains and over time. In this view, an automated system can ultimately only be accurate insofar as it approximates the best available human-coded measure.

Closing Thoughts on Integrative Complexity Metrics

How far is it possible to take the objectification of integrative complexity coding? The short answer is that, as is so often true in life, we never know if we have had enough until we have had more than enough. Our best guess is, however, that current efforts represent an important advance and help to bring into clearer focus the underlying meaning of the theoretical construct and what now needs to be done to measure it in more sophisticated and reliable ways. Conway et al. (2014) are conversation starters, not conversation enders.

ACKNOWLEDGMENTS

We thank H. Andrew Schwartz for his helpful feedback and input about the future of natural language processing. Correspondence concerning this article should be addressed to Philip Tetlock, Department of Psychology, University of Pennsylvania, 3720 Walnut St., Philadelphia, PA 19104-6241. E-mail: tetlock@wharton.upenn.edu

REFERENCES

Baker-Brown, G., Ballard, E. J., Bluck, S., de Vries, B., Suedfeld, P., & Tetlock, P. E. (1992). The conceptual/integrative complexity scoring manual. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 401–418). Cambridge, UK: Cambridge University Press.

Conway, L. G., III, Conway, K. R., Gornick, L. J., & Houck, S. C. (2014). Automated integrative complexity. *Political Psychology, 35*, 603–624.

Conway, L. G., III, Suedfeld, P., & Tetlock, P. E. (2001). Integrative complexity and political decisions that lead to war and peace. In D. J. Christie, R. V. Wagner, & D. Winter (Eds.) *Peace, conflict, and violence: Peace psychology for the 21st century* (pp. 66–75). Englewood Cliffs, NJ: Prentice-Hall.

Conway, L. G., III, Thoemmes, F., Allison, A. M., Towgood, K. H., Wagner, M. J., Davey, K., Salcido, A., Stovall, A. N., Dodds, D. P., Bongard, K., & Conway, K. R. (2008). Two ways to be complex and why they matter: Implications for attitude strength and lying. *Journal of Personality and Social Psychology, 95*(5), 1029–1044.

Crease, R. P. (2011). *World in the balance: The historic quest for a universal system of measurement*. New York, NY: W.W. Norton.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Dickens, C. (2003). *A tale of two cities*. (R. Maxwell, Ed.) London, UK: Penguin Books. (Original work published 1859).

Feist, G. J. (1994). Personality and working style predictors of integrative complexity: A study of scientists' thinking about research and teaching. *Journal of Personality and Social Psychology, 67*(3), 474–484.

Gruenfeld, D. H. (1995). Status, ideology, and integrative complexity on the U.S. Supreme Court: Rethinking the politics of political decision making. *Journal of Personality and Social Psychology, 68*(1), 5–20.

Gruenfeld, D. H., & Hollingshead, A. B. (1993). Sociocognition in work groups: The evolution of group integrative complexity and its relation to task performance. *Small Group Research, 24*, 383–405.

Hermann, M. G. (1999). *Assessing leadership style: A trait analysis*. Columbus, OH: Social Science Automation.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*, 177–196.

Jervis, R. (1976). *Perception and misperception in international politics*. Princeton, NJ: Princeton University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Landauer, T. K, Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284.

Levi, A., & Tetlock, P. E. (1980). A cognitive analysis of Japan's 1941 decision for war. *Journal of Conflict Resolution, 24*(2), 195–211.

Pancer, S. M., Hunsberger, B., Pratt, M. W., Boisvert, S., & Roth, D. (1992). Political roles and the complexity of political rhetoric. *Political Psychology, 13*(1), 31–43.

Porter, C. A., & Suedfeld, P. (1981). Integrative complexity in the correspondence of literary figures: Effects of personal and societal stress. *Journal of Personality and Social Psychology, 40*(2), 321–330.

Satterfield, J. M. (1998). Cognitive-affective states predict military and political aggression and risk taking: A content analysis of Churchill, Hitler, Roosevelt, and Stalin. *Journal of Conflict Resolution, 42*(6), 667–690.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics, 24*(1), 97–123.

Shakespeare, W. (2000). *The tragedy of King Richard II*. (F. E. Dolan, Ed.) New York, NY: Penguin Books.

Suedfeld, P. (1985). APA presidential addresses: The relation of integrative complexity to historical, professional, and personal factors. *Journal of Personality and Social Psychology, 49*(6), 1643–1651.

Suedfeld, P., & Bluck, S. (1988). Changes in integrative complexity prior to surprise attacks. *Journal of Conflict Resolution, 32*(4), 626–635.

Suedfeld, P., & Bluck, S. (1993). Changes in integrative complexity accompanying significant life events: Historical evidence. *Journal of Personality and Social Psychology, 64*(1), 124–130.

Suedfeld, P., Bluck, S., Ballard, E. J., & Baker-Brown, G. (1990). Canadian federal elections: Motive profiles and integrative complexity in political speeches and popular media. *Canadian Journal of Behavioural Science, 22*, 26–36.

Suedfeld, P., de Vries, B., Bluck, S., Wallbaum, A. B. C., & Schmidt, P. W. (1996). Intuitive perceptions of decision-making strategy: Naïve assessors' concepts of integrative complexity. *International Journal of Psychology, 31*(5), 177–190.

Suedfeld, P., & Rank, A. D. (1976). Revolutionary leaders: Long-term success as a function of changes in conceptual complexity. *Journal of Personality and Social Psychology, 34*(2), 169–178.

Suedfeld, P., & Tetlock, P. (1977). Integrative complexity of communications in international crises. *Journal of Conflict Resolution, 21*(1), 169–184.

Suedfeld, P., Tetlock, P. E., & Ramirez, C. (1977). War, peace, and integrative complexity: UN speeches on the Middle East problem, 1947–1976. *Journal of Conflict Resolution, 21*(3), 427–442.

Suedfeld, P., & Tetlock, P. E. (2001). Individual differences in information processing. In A. Tesser & N. Schwartz (Eds.), *The Blackwell handbook of social psychology, Vol. 1: Intraindividual processes* (pp. 284–304). London, UK: Blackwell.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54.

Tetlock, P. E. (1979). Identifying victims of groupthink from public statements of decision makers. *Journal of Personality and Social Psychology, 37*(8), 1314–1324.

Tetlock, P. E. (1981). Personality and isolationism: Content analysis of senatorial speeches. *Journal of Personality and Social Psychology, 41*(4), 737–743.

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45*(1), 74–83.

Tetlock, P. E. (1984). Cognitive style and political belief systems in the British House of Commons. *Journal of Personality and Social Psychology, 46*(2), 365–375.

Tetlock, P. E., & Boettger, R. (1989). Cognitive and rhetorical styles of traditionalist and reformist Soviet politicians: A content analysis study. *Political Psychology, 10*(2), 209–232.

Tetlock, P. E., Hannum, K. A., & Micheletti, P. M. (1984). Stability and change in the complexity of senatorial debate: Testing the cognitive versus rhetorical style hypotheses. *Journal of Personality and Social Psychology, 46*(5), 979–990.

Tetlock, P. E., & Tyler, A. (1996). Churchill's cognitive and rhetorical style: The debates over Nazi intentions and self-government for India. *Political Psychology, 17*(1), 149–170.

Van Hiel, A., & Mervielde, I. (2003). The measurement of cognitive complexity and its relationship with political extremism. *Political Psychology, 24*(4), 781–801.

Wallace, M. D., Suedfeld, P., & Thachuk, K. (1993). Rhetoric of leaders under stress in the Gulf Crisis. *Journal of Conflict Resolution, 37*(1), 94–107.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics, 35*(3), 399–433.

Winter, D. J. (2007). The role of motivation, responsibility, and integrative complexity in crisis escalation: Comparative studies of war and peace crises. *Journal of Personality and Social Psychology, 92*(5), 920–937.

Young, M. D. (2001). Building WorldView(s) with Profiler+. In M. D. West (Ed.) *Progress in communications sciences: Applications of computer content analysis*. (Vol. 17, pp. 17–32). Westport, CT: Ablex.

Young, M. D., & Hermann, M. G. (2014). Integrative complexity has its benefits. *Political Psychology, 35*, 635–645.

Zimmermann, H.-J. (1991). *Fuzzy set theory—and its applications*. (2nd ed.) Boston, MA: Kluwer Academic.