Journal of Social and Political Psychology
jspp.psychopen.eu | 2195-3325

Original Research Reports

Validating Automated Integrative Complexity: Natural Language Processing and the Donald Trump Test

Lucian Gideon Conway III*[a], Kathrene R. Conway[a], Shannon C. Houck[b]

[a] Psychology Department, University of Montana, Missoula, MT, USA. **[b]** Defense Analysis Department, Naval Postgraduate School, Monterey, CA, USA.

Abstract

Computer algorithms that analyze language (natural language processing systems) have seen a great increase in usage recently. While use of these systems to score key constructs in social and political psychology has many advantages, it is also dangerous if we do not fully evaluate the validity of these systems. In the present article, we evaluate a natural language processing system for one particular construct that has implications for solving key societal issues: Integrative complexity. We first review the growing body of evidence for the validity of the Automated Integrative Complexity (AutoIC) method for computer-scoring integrative complexity. We then provide five new validity tests: AutoIC successfully distinguished fourteen classic philosophic works from a large sample of both lay populations and political leaders (Test 1) and further distinguished classic philosophic works from the rhetoric of Donald Trump at higher rates than an alternative system (Test 2). Additionally, AutoIC successfully replicated key findings from the hand-scored IC literature on smoking cessation (Test 3), U.S. Presidents' State of the Union Speeches (Test 4), and the ideology-complexity relationship (Test 5). Taken in total, this large body of evidence not only suggests that AutoIC is a valid system for scoring integrative complexity, but it also reveals important theory-building insights into key issues at the intersection of social and political psychology (health, leadership, and ideology). We close by discussing the broader contributions of the present validity tests to our understanding of issues vital to natural language processing.

Keywords: Automated Integrative Complexity, integrative complexity, natural language processing, AutoIC

Journal of Social and Political Psychology, 2020, Vol. 8(2), 504–524, https://doi.org/10.5964/jspp.v8i2.1307

Received: 2019-08-12. Accepted: 2020-05-14. Published (VoR): 2020-09-02.

Handling Editor: Alessandro Nai, University of Amsterdam, Amsterdam, The Netherlands

*Corresponding author at: 143 Skaggs Building, University of Montana, MT, 59812, USA. E-mail: luke.conway@umontana.edu

> "Applications of automated text analysis measuring topics, ideology, sentiment or even personality are booming in fields like political science and political psychology." Schoonvelde, Schumacher, & Bakker (2019, p. 21)

Research at the intersection of social and political psychology cannot thrive without the continual development of trustworthy and interpretable measurements. As the above quote suggests, it is an exciting time to work in our respective fields, as this era has seen the advent of new, automated methods for scoring long-cherished constructs (e.g., Boyd & Pennebaker, 2017; Schoonvelde, Schumacher, & Bakker, 2019). These new automated methods

– often referred to as *natural language processing* methods – allow researchers to gain new insights by scoring massive amounts of material at levels heretofore unheard of.

However, as the authors also note (Schoonvelde et al., 2019), there is a danger as well: It is possible that we will move ahead with seductively easy-to-score measurements without proper scientific discussion about what they mean – or if they are even measuring what they claim to measure. As a result, ongoing validation of any natural language processing system is vital to the health of the field (see Schoonvelde et al., 2019, for an excellent discussion).

To that end, we evaluate evidence pertaining to the validity of a natural language processing system designed to measure a construct with a long and storied history at the intersection of social and political psychology: *Integrative complexity*.

Human-Scored Integrative Complexity

Designed in its current instantiation by Peter Suedfeld's lab (e.g., Suedfeld et al., 1977), integrative complexity is the measurement of the degree that spoken or written materials have *differentiation* (the recognition of different distinct dimensions) and *integration* (the subsequent recognition of interrelations among differentiated dimensions). At a conceptual level, lower integrative complexity scores indicate rigid, black-and-white communication; higher integrative complexity scores reflect more multi-dimensional language.[i]

Human-scored integrative complexity has proven vitally important in understanding behavior at the intersection of social and political psychology. For example, integrative complexity has been directly tied to the reduction of social problems such as war (Suedfeld & Jhangiani, 2009; Suedfeld, Tetlock, & Ramirez, 1977; Tetlock, 1985; see Conway et al., 2001; Conway et al., 2018, for reviews), terrorism (Andrews Fearon & Boyd-MacMillan, 2016; Conway et al., 2011; Conway & Conway, 2011; Houck et al., 2018), and poor health (Conway et al., 2017). It has further been tied to constructs that are directly related to solving societal problems, such as political ideology (Conway et al., 2016a; Conway et al., 2016b; Houck & Conway, 2019; Suedfeld, 2010; Tetlock, 1983, 1984) and world leaders' success in gaining and keeping power (Conway et al., 2012; Suedfeld & Rank, 1976).

Automated Integrative Complexity

Due to the massive advantages of automated text scoring (see Boyd & Pennebaker, 2017), over the past decade, there has been an increasing push to automate integrative complexity (Conway et al., 2014; Houck et al., 2014; Robertson et al., 2019; Suedfeld & Tetlock, 2014; Tetlock et al., 2014; Young & Hermann, 2014). However, given integrative complexity's strong ties to key societal issues, it is vital that we continuously evaluate such systems. Indeed, the need for further discussion on this topic is highlighted by the fact that, in the last five years, research articles have used the Linguistic Inquiry and Word Count's complexity/analytic thinking score (Pennebaker et al., 2015) as a direct measurement of *integrative complexity* (e.g., Brundidge et al., 2014; Vergani & Bliuc, 2018), in spite of the fact that validation studies show it correlates at only $r = .14$ with expert-scored integrative complexity (Conway et al., 2014).[ii]

In 2014, Conway and colleagues introduced the Automated Integrative Complexity scoring system (*AutoIC*), which, unlike the LIWC, was specifically designed by integrative complexity researchers to measure that construct

(Conway et al., 2014; Houck et al., 2014). The AutoIC system scores differentiation and integration in the same hierarchical fashion as human-scored IC. Although its development was informed by both rudimentary correlational machine learning and expert human input, it was guided far more by human input than by machine learning. Specifically, expert human integrative complexity scorers performed linguistic analysis of every word or phrase that might be associated with integratively complex or integratively simple language, using synonym trees to include as wide a range as possible. After creating an initial system, researchers trained the system on a set of data – using expert human integrative complexity scoring as the benchmark – and then prospectively tested it on an entirely new set of data.

The resulting system has over 3,500 complexity-relevant root words and phrases. Many of these words are lemmatized (e.g., "complex*" scores "complex*ity*," "complex/*y*," etc.) and thus the actual number of scored words/phrases is appreciably higher than the root number. AutoIC breaks documents down into equal-size paragraphs and thus, like human-scored IC, provides paragraph-level averages. The resulting AutoIC algorithm is probabilistic and hierarchical. (1) It is probabilistic because it has 13 separate dictionary categories that differentially assign points to particular words/phrases when they appear, depending on the *probability* that the word/phrase is associated with higher complexity. For example, the phrase "on the other hand" rarely appears without indicating differentiation – so when that phrase appears, 2 points are added (from the base score of 1). (2) Second, AutoIC is hierarchical: Words are parsed into those associated with integration and differentiation in a manner conceptually identical to human scoring. As a result, while it is possible for multiple words/phrases associated with differentiation to add to a score of three, no additional words from differentiation lists would increase the score beyond three. Instead, in a manner conceptually identical to human scoring, achieving an AutoIC score above three requires words from one of several integration lists (words/phrases like "proportional to" and "integrated with").

In the original validity paper, AutoIC (1) showed higher correlations with expert human scorers than other attempts to automate the construct and (2) showed that both the differentiation and integration dictionaries contribute positively to the overall score (Conway et al., 2014). Further, (3) AutoIC replicated effects from human-scored IC in the Bush/Kerry debates, Obama/McCain debates, early Christian writings, and smoking/health domains (Conway et al., 2014).

However, as acknowledged by the authors (Conway et al., 2014; Houck et al., 2014), automating integrative complexity was at that point still in a comparatively early stage of development, and therefore validation evidence for the AutoIC system was in its early stages as well. The purpose of this present paper is to summarize updated evidence for the AutoIC system, provide additional evidence for the system that pertains to key issues in social and political psychology, and discuss the specific contributions AutoIC has made to the social and political psychology literature.

Summary of New AutoIC Validity Evidence to Date

Many different means exist of validating natural language processing systems (for discussions, see Houck et al., 2014; Tetlock et al., 2014; Young & Hermann, 2014).

Overlap With Expert Human Scorers

The most direct kind of validity is the degree that an automated system overlaps with expert human scorers on the same material scored by human coders on a prospective test that was not used in automated "training" (Houck et al., 2014; Tetlock et al., 2014). In the original work (Conway et al., 2014; Houck et al., 2014), AutoIC's average correlation with human scorers at the document level was r = .82. At the paragraph level, the overall correlation was r = .46; for prospective tests only, the paragraph level correlation was r = .41. Since the original paper, several additional studies have also correlated expert human scorers with AutoIC across religious documents (Houck et al., 2018), comparisons of fictional versus non-fictional characters (McCullough & Conway, 2018a), decision-making scenarios (Prinsloo, 2016), and health (Test 3, this paper). As can be seen in Table 1, the correlations with human scorers exceeded the tests from the original validity paper in every case.

Table 1

Direct Validity Evidence: Average Human-AutoIC Correlation

Source	Paragraph Level	Document Level
Conway et al. (2014) Nine Study Ave.	*.46*	*.82*
New Research		
Houck et al (2018)	.46	n/a
McCullough & Conway (2018a)	.48	n/a
Prinsloo (2016) Scenario 1	.47	n/a
Prinsloo (2016) Scenario 2	.52	n/a
Prinsloo (2016) Scenario 3	.53	n/a
Present Article, Test 3	.47	.70

Note. All effect sizes = *r*. Paragraph level = correlations paragraph-by-paragraph. Document level = correlations summarizing the same exact corpus of materials at the appropriate document/person level.

Theoretical Contributions of AutoIC

Additional validity tests involve the existence of theoretically-interpretable findings that were obtained using the measurement tool (Houck et al., 2014). While such findings are not direct evidence of the tool in question's ability to measure complexity per se – because it is possible the findings may have emerged for non-complexity related factors – they are nonetheless important, as any system is in one sense only as good as the interpretable findings it has produced.

In the five years since Conway et al. (2014), evidence showing the descriptive usefulness of the system to understand theoretically-interpretable phenomena has grown. For example, terrorism is a vitally important research area at the intersection of social and political psychology, and yet terrorist groups are very difficult to study. Thus, at-a-distance methods such as integrative complexity have proven to be important in both understanding and preventing terrorism (see, e.g., Conway & Conway, 2011). AutoIC has recently added to our knowledge of terrorism in this regard. Extending prior work using hand-scored IC (Conway et al., 2011; Smith et al., 2008), research has shown that AutoIC is lower in a more extreme terrorist group than a terrorist group using less extreme methods – and that extremity is tied to drops in complexity over time (Houck et al., 2017). Importantly, this work suggests that terrorist groups differ *from other terrorist groups*, such that more violent terrorist groups are lower in complexity (Houck et al., 2017). Other work on terrorists has revealed that peace-based dialogue sessions with convicted

Indonesian terrorists increase terrorists' AutoIC in theoretically-expected ways (Putra et al., 2018). This work with AutoIC suggests that it is possible to use interventions to increase terrorists' integrative complexity in ways consistent with violence-reduction (Putra et al., 2018). Taken together, this work on terrorism has not merely replicated what has come before. Rather, it has revealed new and important avenues for understanding that would not have existed without AutoIC.

AutoIC has similarly advanced our understanding of the individual stability of complex thinking (Conway & Woodard, 2019), the complexity of real versus fictional writings (McCullough & Conway, 2018a), educational interventions (Felts, 2017; Prinsloo, 2016; University of Montana Psychology Department, 2018), the popularity of movies (McCullough & Conway, 2018b), the rated quality of video game dialogue (McCullough, 2019a), the success of fan fiction (McCullough, 2020), critical response to horror films (McCullough, 2019b), and the complexity of twitter (McCullough & Conway, 2019). Thus, AutoIC has begun to offer theoretical insights into multiple important psychological arenas.

Additional Validity Tests: Transitional Summary

In summary, evidence across dozens of studies reveals that (a) AutoIC is moderately correlated with human-scored IC across multiple contexts, and (b) AutoIC helps us understand theoretically-interpretable phenomena across varied domains. However, additional validity tests are needed (see Houck et al., 2014, for discussion). To fill in this gap, we provide 5 new validity tests.

First, Houck et al. (2014) discuss the need for tests that compare groups or conditions on which complexity *ought* to differ. Such tests would intentionally *not* provide carefully controlled conditions attempting to isolate a key variable; rather, they would purposefully stack the proverbial deck, such that it is clear that one group *should* be higher than another group in complexity. Thus, if an automated system failed to distinguish between groups that ought to differ in complexity in this way, it would call its validity into question, in much the same way that Google's object recognition tool was called into question when it was unable to distinguish a cat from an avocado (Ross, 2019). In the present study, we provide two such tests (described in more detail below) – tests that attempt to distinguish higher-complexity groups (classic philosophers) from lower-complexity groups (modern political rhetoric and layperson opinions).

A second valuable piece of validity evidence involves replications of previously-found hand-scored IC effects. In their original paper, Conway et al. (2014) discuss several attempts at such replications. Clearly, however, the area needs more work. In Validity Tests 3-5, we attempt to replicate some of the key findings from three published papers (Conway et al., 2017; Houck & Conway, 2019; Thoemmes & Conway, 2007).

Validity Tests 1 and 2: Comparing Philosophers to Politicians and Lay Populations

We would expect classic philosophical works to be higher in complexity *on average* than political rhetoric or the opinions of lay persons. Classic philosophical works involve some of the greatest minds of all time (thus, those

persons who ought to have the most *ability* to think complexly), with abundant time and cognitive resources (thus, conditions that ought to afford maximum *resources* to think complexly), working through complex problems with the goal of parsing them in a complex way for a highly intelligent audience (thus, *high motivation* and domain-specific likelihood of *complex communication*). As a result, we would expect that classic philosophy should be higher in complexity than modern political speeches designed largely for lay audiences, or the opinions of those lay audiences themselves. Although there are always exceptions, it is nonetheless the case that any complexity system that failed to *consistently* distinguish classic philosophy *on average* from these other forms of communication would be called into question as a complexity-measuring tool (see Houck et al., 2014; Tetlock et al., 2014, for general discussions).

We purposefully selected fourteen of the most famous philosophical works in Western literature: Descartes' *A Discourse on Method*, Plato's *Republic*, Hobbes' *Leviathan*, and Kant's *Critique of Pure Reason*, Berkeley's *A Treatise Concerning the Principles of Human Knowledge*, Epicurus' *Principle Doctrines*, Bacon's *Novum Organum*, Aristotle's *Nicomachean Ethics*, More's *Utopia*, Aquinas' *Summa Theologica (Part I)*, Hobbes' *Leviathan*, Mill's *Utilitarianism,* Russell's *The Problems of Philosophy*, Wittgenstein's *Tractatus Logico*, and Hume's *A Treatise on Human Nature.*

For Validity Test 1, we compare these works' AutoIC scores to modern political rhetoric and lay populations. In Validity Test 2, we do a more focused test comparing these works to one particular modern politician: Donald Trump.

Validity Test 1: Comparing Classic Philosophy to Modern Political Rhetoric and Layperson Opinions

We scored each of the above philosophy works for AutoIC in its entirety. Although for Validity Test 1 we use the philosophy work (N = 14) as the unit of analysis, this scoring of the classic philosophers entailed over 19,000 paragraphs and over 1.4 million words.

For comparison groups in Validity Test 1, we used two groups that we would expect to fall into the average-to-low range for IC: State of the Union speeches from U.S. Presidents and over 37,000 *This I Believe* essays from lay persons.[iii] Both form excellent comparison groups for the present purpose. We would expect SOTU speeches to be low-to-average in complexity: And indeed, when scored by expert human coders, SOTU speeches showed a fairly low mean score (mean IC = 1.78). *This I Believe* essays represent average opinions of typical lay people about what they believe – and thus we would expect them to be lower in complexity *on average* than classic, serious philosophical works.

AutoIC passed this validity test: The AutoIC score for philosophers (M = 2.60) was higher than both the total of U.S. Presidents' State of the Union Speeches, M = 2.10, $F(1, 53) = 78.16$, $p < .001$, and lay persons' "This I Believe" essays, M = 1.96, $F(1, 37446) = 36.47$, $p < .001$. Importantly, converted to r, estimated effect sizes suggested these are very large effects, r's = .77 and .84, p's < .001.[iv]

Validity Test 2: The Donald Trump Test

For Validity Test 2, we compared samples of the classic philosophers discussed above with samples from one specific politician: Donald Trump. We chose Donald Trump because research suggests that Trump is more prone to using simple rhetoric than other politicians (e.g., Ahmadian, Azarshahi, & Paulhus, 2017; Jordan & Pennebaker,

2017), including our own scoring with AutoIC (Conway & Zubrod, 2020). Given that, in general, we would expect classic philosophy to be higher in complexity than political rhetoric for the masses, we ought to *especially* expect classic philosophy to be higher in complexity than a politician for which there are unique reasons to expect low complexity. Thus, this "Donald Trump" test clearly qualifies as a strong expected validity test as described by Houck et al. (2014).

For this additional test, we further compared AutoIC to a new method for scoring integrative complexity: V+POStags (Robertson et al., 2019). Admirably, V+POStags involves both an attempt to create a human-scored vocabulary of words associated with complexity *and* a machine learning approach focused on syntax (see Robertson et al., 2019). Using a different, non-correlational method of evaluation that focuses on discrete categories, Robertson et al. (2019) provided one comparative test (on 30 paragraphs) suggesting V+POStags provides unique benefits (above and beyond AutoIC) for scoring integrative complexity.[v] However, currently, the two systems have not been comprehensively compared. To help fill in this gap, a new and larger validity test needs to be run on both systems without either system having trained on the data set. Below, we provide such a test.

For Validity Test 2, we compared 153 different trials (total number of scored words > 300,000, AutoIC paragraphs > 4,000) for AutoIC and V+POStags on the exact same materials. Specifically, we scored all of Donald's Trump's 2016 presidential debates with Hillary Clinton for AutoIC and V+POStags.[vi] (All words not from Donald Trump, e.g., the moderator or Clinton, were removed prior to scoring). In order to standardize the sample length across materials, we broke up the philosophical works scored in Test 1 into groupings approximately equal to the average word length of Trump's discussions in the three debates (around 7,400 words).[vii] We then randomly selected four groupings from each philosophic work for scoring both by AutoIC and V+POStags (most works had a least four groupings; for works that did not have 4 groupings, we created as many as the material allowed).

Our primary question of interest is the degree that AutoIC and V+POStags can each consistently distinguish Trump from classic philosophy. To accomplish this, we compared (separately for each system) each debate transcript score against each philosophical grouping score. This provided 153 separate comparisons to evaluate if each system assigned a higher score to a famous philosophical work than to Donald Trump. When a comparison yielded a higher score for classic philosophy, it was counted a *successful* trial; when the comparison did not yield a higher score for classic philosophy, it was counted as a *failed* trial. It is worth noting that throughout, AutoIC and V+POStags scored the *exact same materials*.

AutoIC assigned a higher complexity score to the philosophic work on all 153 trials (100% success). However, V+POStags assigned a higher score to the philosophic work on 33% (51 of 153) of the trials.

Discussion of Validity Tests 1 and 2

Validity tests that demonstrate expected differences between groups on a linguistic variable are vital forms of natural language processing validity (Houck et al., 2014), and yet no such validity tests currently exist for automated integrative complexity measurements. Validity Tests 1 and 2 help fill in this gap. Across both tests, AutoIC consistently showed higher levels of complexity for classic philosophical works than for politicians and laypersons.

Validity Test 2 further compared AutoIC to a newly-developed system (V+POStags). This test revealed that while AutoIC showed 100% success on this validity test, V+POStags showed comparatively less success (33%). Why might this be? We suspect there are two independent reasons. First, there is a tradeoff between machine learning

and human learning in system development. Machine learning is excellent at detecting patterns in large datasets that humans cannot detect. However, it is less good at predicting alterations that might occur to those patterns in new data that it was not "trained" on. While both AutoIC and V+POStags used both human learning and machine learning in development, AutoIC focused mostly on human learning and V+POStags focused comparatively more on machine learning. Thus, one possible reason for AutoIC's success is that human-learning developed enterprises are more stable across contexts. We return to this larger issue in the discussion.

We also suspect that part of the reason AutoIC outperformed V+POStags on this validity test is more specific to the V+POStags methodology (and not machine learning in general). Specifically, a closer look at the numbers for V+POStags suggests that part of the problem is the commitment of that system to assigning discreet integers, instead of scoring (as AutoIC does) on a sliding scale. Indeed, V+POStags assigned the exact same score to all philosophic works (IC = 3), improbably suggesting there is no variability among the philosophers on IC. Further, the additional probability assessments provided by V+POStags for each discrete score (1, 2, 3, etc.) suggest that there is a strong tendency for the philosophers (compared to Trump) to be assigned higher probability values associated with higher scores – a fact that suggests the system is discarding useful variability in assigning scores.

Of course, while the Trump test is an important step in our understanding of natural language processing validity, we do not want to make *too* much of one validity test comparison. Nor are we suggesting that V+POStags has no value. Quite the contrary: We believe the V+POStags system is an excellent (and much-needed) machine learning-focused effort for the natural language processing of integrative complexity, and we commend the authors for their work in this regard. Rather, our evaluation of this test is that V+POStags is a promising system that, like all newly-developed systems, requires more work to fulfill that potential.

Validity Tests 3-5: Replication of Existing Studies

Below, we further provide three additional validity tests of a different type. Using AutoIC, we attempt to replicate key aspects of three published studies that originally used human-scored IC. Table 2 provides a larger summary of these data. As can be seen there, we use a two-fold rubric for evaluating these replication attempts. (1) First, computing similar tests as the original studies, we evaluate whether or not the replication attempt showed a similarly-significant result in the same direction as the original study. We consider a *successful replication* in this regard if the original study showed a significant effect that is also significant in the replication, or if the original study showed no significant effect and the replication attempt identically showed a non-effect. (2) However, we further compute common effect size metrics for each study and compare the *strength* of each effect (significant or otherwise) for each comparable effect for the original study and the AutoIC replication attempt. As can be seen in Table 2, we provide not only these descriptive statistics for each comparison, but also a brief subjective summary comment for ease of discussion.

Below, we briefly describe each test and offer a narrative summary of the outcomes.

Validity Test 3: Replication of Smoking Attitudes Study From Conway et al. (2017)

Few problems are more pressing in modern society than the issue of health – and smoking remains one of the largest health issues in the world (see Conway et al., 2017). Conway et al. (2017) scored Motivational Interviewing counselling sessions from a prior grant-funded study on smoking cessation (see Harris et al., 2010, for details)

for hand-scored integrative complexity. The primary finding of Conway et al. (2017) was that both counselors and clients in sessions that ultimately led to quitting smoking (Successful Quitters) showed lower integrative complexity during four sessions than in sessions for clients who attempted to quit and failed (human-scored d = 0.74; for clients only, d = 0.76; for counsellors only, d = 0.72). Conway et al. (2017) scored five paragraphs per session (total paragraph N = 1100).

Table 2

Replication Comparison Effect Sizes and Inferential Statistics from Validity Tests 3-5

Source	Human-Scored Original Study	Auto IC Replication	Summary
Test 3, Whole Corpus			
IC reduces smoking (all)	*0.74***	*0.39***	*Successful replication w/smaller effect size*
IC reduces smoking (clients only)	0.76***	0.56*	Successful replication w/smaller effect size
IC reduces smoking (couns. only)	0.72***	0.76***	Successful replication
Test 3, Identical Corpus			
IC reduces smoking (all)	*0.74***	*0.77***	*Successful replication*
IC reduces smoking (clients only)	0.76***	0.60**	Successful replication w/smaller effect size
IC reduces smoking (couns. only)	0.72***	0.95***	Successful replication
Test 4, Year of Term Effects			
Year 1-Year 4 Reduction	*0.33***	*0.06**	*Successful replication w/smaller effect size*
Year X Win/Loss Interaction	.sig*	ns	Failed replication
Test 4, Personality Effects			
Overall IC Personality Variance	*0.37***	*0.80***	*Successful replication w/larger effect size*
Affiliation-IC correlation	0.40*	0.35*	Successful replication
Liberalism-IC correlation	0.23	0.19	Successful replication
Friendliness-IC correlation	0.32	0.12	Pattern replication w/smaller effect size
Wittiness-IC correlation	0.42*	0.24	Pattern replication w/smaller effect size
Inflexibility-IC correlation	-0.31*	-0.08	Pattern replication w/smaller effect size
Extraversion-IC correlation	0.36*	-0.08	Failed replication
Brilliance-IC correlation	0.08	0.24	Pattern replication w/larger effect size
Test 5			
IC-Conservatism, Public Figures	*-0.37**	*-0.12**	*Successful replication w/smaller effect size*
IC-Conservatism, Laypersons	-0.01	-0.03	Successful replication

Note. Effect sizes for Test 4 Year-of-Term effects are *d*'s. Effect size for Overall IC Personality Variance score is a One-Way Intraclass Correlation Coefficient (ICC). All other effect sizes are *r*'s.

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

In validity Test 3, we first scored *all* the materials for each session (total paragraph N = 6,906). As can be seen in Table 2, the primary findings obtained with hand-scoring replicated directly with AutoIC: Sessions with successful quitters showed lower levels of AutoIC overall than failed attempters, $F(1, 217) = 17.32$, d = 0.39, p < .001. This effect also held, as in the original study, for clients, $F(1, 107) = 3.97$, d = 0.56, p = .049, and counsellors, $F(1, 108) = 15.60$, d = 0.76, p < .001, separately.

Inferentially identical results emerged for Validity Test 3 if we only used the 1100 paragraphs from the original Conway et al. (2017) study, with AutoIC showing significantly lower complexity for successful quitters than failed

attempters, $F(1, 217) = 32.00$, $d = 0.77$, $p < .001$. This effect on the identical corpus from the original study also similarly held for clients, $F(1, 107) = 9.74$, $d = 0.60$, $p = .002$, and counsellors, $F(1, 108) = 24.37$, $d = 0.95$, $p < .001$. separately. Although no inferential differences emerged between the whole corpus and the identical corpus, narrowing the focus to only those materials scored in the original study did yield effect sizes that were more in the range of the original study (see Table 2).

These results importantly contribute to our understanding of the relationship between smoking behavior and complexity during counselling sessions. Contrary to the assumption that complexity is an unqualified panacea, often complexity in health contexts can backfire because people need simple-minded focus to make positive health-related change (see Conway et al., 2017). Yet, despite the potential utility of this idea, data testing the effects of complexity is health contexts is scarce. Thus, the present results importantly validate this original finding. And, because (unlike the original study), this study scored the entire corpus of materials, they rule out the possibility that something about the original selection process may have influenced the results. Further, because AutoIC is much faster than hand-scoring, validating AutoIC for this context opens up a tool for researchers that might be very pragmatically and theoretically useful moving forward.

Validity Test 4: Replication of U.S. Presidents' Study From Thoemmes and Conway (2007)

Thoemmes and Conway (2007) hand-scored IC for all first-term U.S. Presidents' State of the Union (SOTU) speeches (through G. W. Bush). They scored up to five paragraphs per speech (679 paragraphs total). For AutoIC, we scored *all* the materials for each SOTU speech (paragraph $N = 18{,}495$) for first-term U.S. Presidents through G. W. Bush. Thus, we drew from the same corpus of materials as the original paper, but (unlike the original paper) we did not randomly sample paragraphs, instead scoring all the materials in that corpus.[viii]

We attempt to replicate findings from the Thoemmes and Conway (2007) study that fall into two categories: (1) Whether or not patterns systematically differed over four years in office, and (2) whether or not individual differences across presidents were in evidence.

Systematic Patterns Over Time

The primary large-scale finding of Thoemmes and Conway (2007) using hand-scored integrative complexity was that SOTU speeches tended to drop for all presidents over the course of the first term. This primary result was replicated with AutoIC: There was a similar main effect of Year of Speech, $F(3, 18492) = 4.65$, $p = .003$. Descriptive results for this pattern were very close to those using hand-scoring reported by Thoemmes and Conway (2007): There was a drop in complexity from year 1 to year 4. Consistent with other work using large samples, the effect sizes were smaller for these comparisons using AutoIC than in the original study (Averaging Year 1-4 and Year 2-4 comparisons, Original Study $d = 0.33$, AutoIC effect = .06). Overall, however, the AutoIC *pattern* closely (and significantly) replicates that of Thoemmes and Conway (2007).

Thoemmes and Conway also reported an interaction between success and year of term for hand-scored IC. Unlike in Thoemmes and Conway (2007), for AutoIC there was no significant interaction between success and year of term for complexity ($p = .629$), and the directional pattern bore little resemblance to the one from the original study.

Taken together, what are we to make of these results? First, they importantly reaffirm one of the basic conclusions of the Thoemmes and Conway (2007) article: That complexity of U.S. presidents tends to drop over their four SOTU speeches. Because of the practical and theoretical significance of this finding, a replication of it not only

validates AutoIC, but simultaneously provides needed triangulating support to the effect of time in office on integrative complexity.

Why did the time in office X electoral success interaction not replicate using AutoIC? There are several possibilities. (1) It is possible that human scoring is tracking a nuance that is important in the existence of the effect – a nuance that AutoIC does not score as effectively. (2) Of course, a failure to replicate can occur for multiple reasons that have little to do with the system under scrutiny (see, e.g., Conway et al., 2014). For example, it is conceptually possible that, because AutoIC is scoring vastly more of the SOTU materials, this failure to replicate casts doubt on the original finding (perhaps if the original study had scored the other 96.3% of the material, it would have likewise showed a non-effect in this case). This potential itself is an important contribution. We cannot completely know the exact cause of a failure to replicate in the present case without more data – but, importantly, in addition to validating the original drop-over-time finding of the Thoemmes and Conway (2007) study, the present results suggest that larger-scale election studies are needed understand the relationship between electoral success and integrative complexity.

Individual Differences

Individual differences-based tests were also provided by Thoemmes and Conway (2007). Importantly, replicating Thoemmes and Conway (2007), the present results showed an effect of the individual president, $F(40, 18456) = 9.00$, $ICC = .80$, $p < .001$, suggesting that part of the variance in complexity is accounted for by individual differences between persons.

Thoemmes and Conway also attempted to ascertain what personality traits might be associated with presidential complexity by correlating trait scores for each president with their overall IC score. We used AutoIC to perform identical analyses with these personality traits. These results are presented in full in Table 2. Generally, these analyses reveal a similar pattern of results for AutoIC as for human-scored IC, although the AutoIC pattern is weaker overall (average effect size for human-scored IC = .30; for AutoIC = .16). Given that one of the best predictors of integrative complexity has generally been affiliation-related variables (see, e.g., Thoemmes & Conway, 2007), it is perhaps noteworthy AutoIC (like human-scored IC) showed a significant positive correlation with affiliation motive ($r = .35$, $p = .049$). AutoIC also showed a positive relationship with political liberalism that is not only almost identical to that used in the original Thoemmes and Conway (2007) study, but is further validated across multiple studies of politicians via meta-analyses (Houck & Conway, 2019).

Taken together, these results provide an important contribution to our understanding of presidential integrative complexity. First, the original Thoemmes and Conway finding that substantive variance is attributable to stable differences across presidents has been discussed as one of the few empirical investigations into individual differences in politicians' complexity (see Conway & Woodard, 2019). Given the vital implications of understanding the degree that persons are (or are not) chronically complex, the present replication's finding that individual variability in presidential integrative complexity accounts for a significant percentage of the variance is important. It further validates additional recent work (Conway & Woodard, 2019) suggesting that integrative complexity can reasonably be construed, in part, as an individual difference variable.

The present results also generally validate the conclusions of Thoemmes and Conway (2007) concerning what the chronically complex person might look like. That person is especially likely to be high in the affiliation motive

and (to a lesser degree) liberal. While it is tempting to over-interpret differences across the studies, it seems clear that, in the main, these results tend to point to roughly similar conclusions as the original study.[ix]

Validity Test 5: Replication of Meta-Analysis on Political Ideology From Houck and Conway (2019)

Some prior work suggests that liberals use more complex rhetoric than conservatives (e.g., Tetlock, 1983, 1984, 1985; see Jost et al., 2003, for a summary), while other work suggests no differences between liberals and conservatives in their use of complex rhetoric (e.g., Conway et al., 2016a; see Houck & Conway, 2019, for a summary). To help resolve this puzzle, Houck and Conway (2019) performed a meta-analysis of 35 studies that had measurements of integrative complexity and political ideology. Because this test used only precise measurements of both constructs – for example, they only used political ideology measurements that were unlikely to be contaminated with complexity-relevant variables such as dogmatism or authoritarianism – this study provides a litmus test of the relationship between ideology and the use of complex language.

Houck and Conway's (2019) results suggested a clear resolution to the puzzle of the ideology-complexity relationship: Whereas liberal political elites were significantly more complex than their conservative counterparts, liberal and conservative laypersons showed very similar levels of complexity. Drawing on previous work in other domains on strategic ideological communication (Conway et al., 2017; Repke et al., 2018; Tetlock, 1981), Houck and Conway (2019) suggested this difference is due to differing norms for conservatives and liberals that cause liberal (but not conservative) politicians to strategically alter their communications to better meet the expected norms of their populaces.

Houck and Conway's (2019) meta-analysis is one important piece of evidence in our understanding of the ideology-complexity relationship. However, it is increasingly important to provide multiple triangulating tests of a particular theory or model (see Crandall & Sherman, 2016), especially when the issue is as hotly debated as the ideology-complexity link (e.g., Baron & Jost, 2019; Clark & Winegard, 2020). The present study provides a conceptual replication of Houck and Conway's (2019) model on an almost entirely new set of data, using AutoIC (as opposed to hand-scored IC) for measuring integrative complexity.

In the present study, we performed a mini meta-analysis (see Goh, Hall, & Rosenthal, 2016) on samples of data collected and scored for AutoIC by the authors. From this potential sample, we followed the same inclusion criteria, coding procedures, and analytic strategy as used in Houck and Conway (2019), except we exclusively incorporated AutoIC studies (and not hand-scored IC). Please see the Supplementary Materials for more detailed descriptions of the samples and criteria.

Results are presented in Table 3 and Table 4. Consistent with a strategic ideological communication model and Houck and Conway (2019), the public nature of the sample moderated the effect of political conservatism on integrative complexity: whereas in samples with public political officials, conservatives were significantly less complex than liberals, r = -.12, p = .043, 95% CI [-.23, -.00], this effect did not emerge in private layperson samples, r = -.03, p = .184, 95% CI [-.06, +.01].

Table 3

Validity Test 5: Correlations Between Conservatism and Integrative Complexity in Public Political Figures and Private Lay Citizen Samples

Sample Characteristic	Conservatism Measure Used	Materials Scored	*N* (speaker)	*n* (speech/doc)	Effect Size Est. (*r*)
Public Political Figures					
Conway & Zubrod (2020)	Ideology ratings	SOTU Speeches	40	147	-.09
Conway & Zubrod (2020)	Party identification	Presidential debates	24	62	-.09
Conway & Zubrod (2020)	Party identification	2016 primary debates	21	142	-.16
Mean Effect Size					**-.12***
Private Citizens					
Conway et al. (unpub.)					
Topic 1	Self-reported ideology	Rep. Leader Stem	202	202	+.02
Topic 2	Self-reported ideology	Dem. Leader Stem	202	202	+.01
Conway et al. (unpub.)					
Topic 1	Self-reported ideology	Smoking Att. Stem	4764	4764	-.06
Topic 2	Self-reported ideology	Pol. Cooperation Stem	4764	4764	-.08
Conway et al. (2014)	Self-reported ideology	Pol. and Social Stems	325	325	-.02
Crawford et al. (unpub.)					
Topic 1	Self-reported ideology	Prayer Law Stem	249	249	+.10
Topic 2	Self-reported ideology	Protest Rights Stem	501	501	-.02
Topic 3	Self-reported ideology	President Stem	501	501	+.02
Mean Effect Size					**-.03**

Note. *N* = number of speakers is the unit of analyses from which the results are based; *n* = number of documents/speeches. The correlation for SOTU speeches is different from that reported for Validity Test 4 because each study required a different unit of analysis to match the different criteria used in each parent study.

Table 4

Validity Test 5: Public/Private Sample as a Moderator of the Political Conservatism and Integrative Complexity Relationship

Moderator	*k*	*r*	*p*	95% CI	Q_b	Q_w
Total Set	11	-.03^	.058	[-.07, +.00]		9.53
Public Versus Private					2.28	
Public Official	3	-.12*	.043	[-.23, -.00]		0.35
Private Citizen	8	-.03	.184	[-.06, +.01]		9.18

Note. *k* = number of studies; *r* = effect size estimate, Pearson's *r*; CI = confidence interval. Q_b = homogeneity between groups; $Q_{w=}$ homogeneity within groups.

^$p < .07$. *$p < .05$.

These results provide important additional evidence that the relationship between political conservatism and complexity differs for public officials and private citizens. In 11 separate samples (encompassing 5,877 persons, 11,859 documents, and 40,428 paragraphs), political conservatives showed a significant negative relationship to integrative complexity for public political officials, but no such relationship for private citizens. This basic pattern is identical to that of a separate meta-analysis (Houck & Conway, 2019) using hand-scored IC on a largely non-overlapping set of materials. Given the current ongoing controversy in political and social psychology concerning the symmetry of "rigidness" measures across conservatives and liberals (e.g., Baron & Jost, 2019; Clark &

Winegard, 2020), evidence validating a clear moderating variable for the ideology-rigidity relationship is vital. The present 11-study meta-analysis provides one such piece of triangulating evidence.

General Discussion: Implications of These Results for Natural Language Processing in the Social and Political Sciences

Five validity tests provide further evidence that AutoIC is a valid measurement of human-scored integrative complexity. Tests 1 and 2 revealed that AutoIC consistently distinguished high- and low-complexity groups/individuals from each other. Tests 3-5 provided replications of key effects from the human-scored IC literature. Taken together with the existing literature, this set of results provides a large array of evidence that AutoIC is effectively measuring human-scored IC. Below, we discuss broader implications of this work for our understanding of natural language processing of complexity in the political and social sciences.

Effect Sizes of Natural Language Processing and Big Data

Although in most cases the basic pattern and inferential statistics were identical to prior studies for Tests 3-5, the replication attempts using AutoIC generally yielded smaller effect sizes. What does this mean?

There are two potential reasons why AutoIC yielded smaller effect sizes, on average, than human-scored IC. (1) Others have noted that big data – whether natural language processing data or otherwise – can often produce smaller effect sizes (see, e.g., Slavin & Smith, 2009; see also Houck et al., 2018; Kramer, Guillory, & Hancock, 2014). Thus, it is possible that the effect sizes for AutoIC are smaller than human-scored IC simply because it generally scores a much larger amount of material. (2) It is also possible that effect sizes differed for a simpler reason: The present AutoIC results generally did not draw from the exact same paragraphs as the human-scored studies. AutoIC and human-scored IC results might be more similar if we had more frequently been able to use the exact same paragraphs for both systems.

The present results cannot definitively distinguish between these two possibilities, but they can offer some clues. In Validation Test 3, using an identical corpus showed effects more similar in size to human-scored IC than those using the whole corpus (see Table 2, Test 3 Whole Corpus versus Identical Corpus; note that Test 3 had both Whole and Identical Corpuses). However, both sets of analyses used the same aggregated unit (meaning they had the same *n* for computing effect sizes). This suggests that the increased AutoIC effect size for the identical corpus was not due to a *general* "large *n*" problem, but rather to AutoIC scoring the same exact set of materials as the original (and thus giving it the best chance at replication due to direct overlap). The paragraph-by-paragraph match hypothesis is further bolstered by the fact that in the present work we found smaller AutoIC effect sizes for the personality measurements for Test 4 (where an identical unit of analysis was used, but no paragraph-by-paragraph match). Taken together, this set of results might argue that the lower effect sizes in this work are generally due to the lack of specificity (and not to a general large-*n* problem), and if we had the exact same materials available for scoring, AutoIC effect sizes would be closer to their human-scored counterparts.

However, because we have few cases that can distinguish between various competing explanations, it is still possible that that a more general *big data = small effect size* problem might account for some of our smaller effect sizes. Importantly, while others have commented on effect size issues with big data/natural language processing and argued that we should not dismiss subsequent small effect sizes out of hand (see, e.g., Slavin & Smith, 2009; see also Houck et al., 2018; Kramer, Guillory, & Hancock, 2014), no study that we know of attempts to set empirical boundaries on the exact parameters of when (and how much) the big data effect size reduction occurs, or (conversely) at what point additional data becomes redundant in natural language processing (e.g., Schönbrodt & Perugini, 2013). Moving forward, it would be very useful for social scientists to more fully explore this issue in empirical studies.

Machine Learning Versus Human Learning

Why does she get a blaster and I don't?...Would you like to know the probability of her using it against you? It's high…It's very high. (K-2SO, from the movie Rogue One: A Star Wars Story)

When artificial intelligence android K-2SO claims that the probability of Jyn Erso betraying them is "very high" in the *Star Wars* movie *Rogue One*, it carries a lot of weight to the intended audience. After all, we tend to view computer intelligences as unburdened with human limitations such as slow processing speed and emotional biases. Similarly, there may be a tendency to assume that "machine learning" is a superior method of approaching any linguistic problem. However, the truth is that machine learning has clear positive and negative trade-offs. Indeed, when we originally laid plans for an automated integrative complexity system, we first used a rudimentary machine learning approach by evaluating which words and punctuations were associated with higher or lower complexity scores. What we found was that, while we could construct effective algorithms for each data set this way, what worked in one data set often failed on another (see Houck et al., 2014).

The reason for this is clear: No matter how sophisticated the algorithm, a computer algorithm is entirely based on what happened in the specific corpus of materials under study. A human, however, has two advantages that a computer does not have. (1) A human has a much, *much* larger corpus of materials to work with when approaching any problem. In other words, a human is not constrained by the body of materials under investigation, because humans have had a lifetime of experience with the language under study *in general*. (2) Partially as a result of this, a human can imagine what would potentially happen in other scenarios that the data in question do not remotely cover. For example, imagine that in one dataset the term *"on the other hand"* ("Republicans' foreign policy is bad; *on the other hand*, their economic policy is…") and the term *"apart from"* ("quite *apart from* the influence of the war in Iraq, Bush's domestic policies…") are consistently used by one political author to mean clear differentiation (IC score of 3). Based on these data, a computer algorithm would subsequently assign both of those phrases a high probability score for 3. But human coders would do something else entirely. They would realize that, in many other contexts, the phrase "apart from" is actually used in a purely descriptive fashion that implies no complexity at all ("I do not wish to be *apart from* you"), whereas the number of contexts that "on the other hand" is likely to be used in a non-complex way is comparatively much smaller. Therefore a human-based approach would use the data from the computer learning in a different way – to make estimates of how each phrase would fare *beyond the dataset in question* – and thus assign "on the other hand" a higher complexity probability score than "apart from."

This is one of the reasons that, we believe, AutoIC has consistently performed at similar levels across multiple new contexts beyond those it was originally designed on, and systems such as V+POStags (which focus more

extensively on machine learning) have done more poorly when faced with a new context (such as the philosophic writings from the current paper). Indeed, AutoIC researchers spent a larger proportion of time developing human-generated dictionaries of words and phrases than other researchers have done. Consider that, in contrast to V+POStags – which spent more time on machine learning and thus developed a human dictionary that had 312 base words – AutoIC has thousands of complexity-related words and phrases in its lexicon. Further, AutoIC researchers spent more time estimating the probability of each word or phrase's contribution to complexity, whereas V+POStags researchers used a simple binary classification system (see Robertson et al., 2019) that lost potential human-inspired nuance.

Of course, machine learning has advantages too – machine learning can often uncover complex relationships that humans cannot. We expect that, as advances in machine learning grow, it will be used more and more effectively. Indeed, one of the clear implications of the above line of reasoning is that the greatest current need for machine learning approaches is a *larger corpus of human-scored IC data* for machine-learning development. The best way to deal with the problem of continuity across datasets head-on is to use as large and as varied a set of data as possible to develop natural language processing systems on. Because human-scoring of IC takes a lot of time, this corpus is currently not large enough to likely be sufficient. However, an important goal of future research should be to expand that corpus so that it is large enough to more fully take advantage of the strengths of machine learning.

Thus, our point is not to undermine the validity of machine learning, but simply to point out that is has both great strengths *and* severe limitations, and to encourage methods-building from multiple perspectives based on rigorous scientific standards (see Schoonvelde, Schumacher, & Bakker, 2019, for a discussion). While ultimately the next generation of improvements will likely indeed come from machine learning – and we would applaud those improvements – we should not assume that just because a system was developed via "machine learning" that it is *de facto* an improvement. This is an empirical field, and those assumptions (however appealing) must still be put to the empirical test.

Notes

i) Please see the Supplementary Materials for specifics about the scale and measurement examples.

ii) Evidence suggests that LIWC's measurements are relevant to complex thinking/rhetoric (see, e.g., Boyd et al., 2015; Jordan et al., 2019), but not to integrative complexity specifically. We guess that researchers are simply unaware that the measure they view as a measurement of "integrative complexity" is largely unrelated to that instantiation of complexity.

iii) *This I Believe* essays were first scored for AutoIC in Houck et al. (2018). However, their use in the present research is entirely novel.

iv) To compute the effect size for the *This I Believe* dataset, we randomly selected a number of participants equal to the number of philosophers and used that randomly selected set to compute the reported effect size.

v) They further provide a validity test showing that V+POStags finds an expected effect in a social media analysis. However, they did not score AutoIC on this subsequent test.

vi) These debates were originally scored for a different project (Conway & Zubrod, 2020), and we were aware of the *general* AutoIC scores for both Trump and the philosophers before running this test.

vii) In order to be sure we were using the V+POStags system as intended by its authors, we consulted with the primary author of the V+POStags paper about issues pertaining to sample length and system use prior to scoring.

viii) The original Thoemmes and Conway (2007) paper used a hard copy (and not electronic) based sampling system, and thus no electronic paragraphs are available for computer scoring from that original work. As a result, we did not (as in Test 3) also score the exact same corpus of paragraphs for Test 4.

ix) For both Validity Test 3 and Validity Test 4, we also computed correlations between AutoIC and human-scored IC. Supplementary Materials report these correlations in full. Of note is that the only correlations available on the exact same materials (n = 1100) were similar to the average reported in the original validity study (paragraph-level r = .47; session-level r = .70).

Funding

Portions of this work were supported by the National Cancer Institute at the National Institutes of Health [Grant Number R15CA186247; Conway PI] and the National Institute of General Medical Sciences, Clinical and Translational Research Infrastructure Network Pilot Grant Program [Grant Number 14-746Q-UMT-PG2-00, Conway PI].

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

The authors have no support to report.

Supplementary Materials

The supplementary materials provide additional tests of human-computer correlations for Validity Tests 3 and 4, more detail about the methods and samples for Validity Test 5, and examples of integrative complexity scoring (for unrestricted access, see Index of Supplementary Materials below).

Index of Supplementary Materials

Conway, L. G., III, Conway, K. R., & Houck, S. C. (2020). *Supplementary materials to "Validating Automated Integrative Complexity: Natural language processing and the Donald Trump Test"* [Additional tests, information, and examples]. PsychOpen. https://doi.org/10.23668/psycharchives.3359

References

Ahmadian, S., Azarshahi, S., & Paulhus, D. L. (2017). Explaining Donald Trump via communication style: Grandiosity, informality, and dynamism. *Personality and Individual Differences, 107*, 49-53. https://doi.org/10.1016/j.paid.2016.11.018

Andrews Fearon, P., & Boyd-MacMillan, E. M. (2016). Complexity under stress: Integrative approaches to overdetermined vulnerabilities. *Journal of Strategic Security, 9*, 11-31. https://doi.org/10.5038/1944-0472.9.4.1557

Baron, J., & Jost, J. T. (2019). False equivalence: Are liberals and conservatives in the United States equally biased? *Perspectives on Psychological Science, 14*(2), 292-303. https://doi.org/10.1177/1745691618788876

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences, 18*, 63-68. https://doi.org/10.1016/j.cobeha.2017.07.017

Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., & Mihalcea, R. (2015, April). Values in words: Using language to evaluate and understand personal values. In *Ninth International AAAI Conference on Web and Social Media* (pp. 31-40). Palo Alto, CA, USA: The AAAI Press.

Brundidge, J., Reid, S. A., Choi, S., & Muddiman, A. (2014). The "deliberative digital divide:" Opinion leadership and integrative complexity in the US political blogosphere. *Political Psychology, 35*(6), 741-755. https://doi.org/10.1111/pops.12201

Clark, C. J., & Winegard, B. M. (2020). Tribalism in war and peace: The nature and evolution of ideological epistemology and its significance for modern social science. *Psychological Inquiry, 31*(1), 1-22. https://doi.org/10.1080/1047840X.2020.1721233

Conway, L. G., III, & Conway, K. R. (2011). The terrorist rhetorical style and its consequences for understanding terrorist violence. *Dynamics of Asymmetric Conflict, 4*, 175-192. https://doi.org/10.1080/17467586.2011.627940

Conway, L. G., III, Conway, K. R., Gornick, L. J., & Houck, S. C. (2014). Automated integrative complexity. *Political Psychology, 35*, 603-624. https://doi.org/10.1111/pops.12021

Conway, L. G., III, Gornick, L. J., Burfiend, C., Mandella, P., Kuenzli, A., Houck, S. C., & Fullerton, D. T. (2012). Does simple rhetoric win elections? An integrative complexity analysis of U.S. presidential campaigns. *Political Psychology, 33*, 599-618. https://doi.org/10.1111/j.1467-9221.2012.00910.x

Conway, L. G., III, Gornick, L. J., Houck, S. C., Anderson, C., Stockert, J., Sessoms, D., & McCue, K. (2016a). Are conservatives really more simple-minded than liberals? The domain specificity of complex thinking. *Political Psychology, 37*, 777-798. https://doi.org/10.1111/pops.12304

Conway, L. G., III, Gornick, L. J., Houck, S. C., Hands Towgood, K., & Conway, K. R. (2011). The hidden implications of radical group rhetoric: Integrative complexity and terrorism. *Dynamics of Asymmetric Conflict, 4*, 155-165. https://doi.org/10.1080/17467586.2011.627938

Conway, L. G., III, Harris, K. J., Catley, D., Gornick, L. J., Conway, K. R., Repke, M. A., & Houck, S. C. (2017). Cognitive complexity of clients and counselors during motivation-based treatment for smoking cessation: An observational study on occasional smokers in a U.S. college sample. *BMJ Open, 7*, Article e015849. https://doi.org/10.1136/bmjopen-2017-015849

Conway, L. G., III, Houck, S. C., Gornick, L. J., & Repke, M. A. (2016b). Ideologically-motivated perceptions of complexity: Believing those who agree with you are more complex than they are. *Journal of Language and Social Psychology, 35*, 708-718. https://doi.org/10.1177/0261927X16634370

Conway, L. G., III, Suedfeld, P., & Tetlock, P. E. (2001). Integrative complexity and political decisions that lead to war or 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, USA: Prentice-Hall.

Conway, L. G., III, Suedfeld, P., & Tetlock, P. E. (2018). Integrative complexity in politics. In A. Mintz (Ed.), *Oxford handbook of behavioral political science*. Oxford, United Kingdom: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190634131.013.7

Conway, L. G., III, & Woodard, S. R. (2019). Integrative complexity across domains and across time: Evidence from political and health domains. *Personality and Individual Differences, 155*, Article 109713. https://doi.org/10.1016/j.paid.2019.109713

Conway, L. G., III, & Zubrod, A. (2020). *The integrative complexity of Donald Trump: Is Trump a unique outlier or an extension of a republican trend towards simplicity?* Manuscript in progress.

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology, 66*, 93-99. https://doi.org/10.1016/j.jesp.2015.10.002

Felts, N. A. (2017). *Please explain yourself: Mechanisms of opinion improvement in deliberative forums* (Doctoral dissertation, The Ohio State University, Columbus, OH, USA). Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=osu1492555075881149

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*, 535-549. https://doi.org/10.1111/spc3.12267

Harris, K. J., Catley, D., Good, G. E., Cronk, N. J., Harrar, S., & Williams, K. B. (2010). Motivational interviewing for smoking cessation in college students: A group randomized controlled trial. *Preventive Medicine, 51*, 387-393. https://doi.org/10.1016/j.ypmed.2010.08.018

Houck, S. C., & Conway, L. G., III. (2019). Strategic communication and the integrative complexity-ideology relationship: Meta-analytic findings reveal differences between public politicians and private citizens in their use of simple rhetoric. *Political Psychology, 40*(5), 1119-1141. https://doi.org/10.1111/pops.12583

Houck, S. C., Conway, L. G., III, & Gornick, L. J. (2014). Automated integrative complexity: Current challenges and future directions. *Political Psychology, 35*, 647-659. https://doi.org/10.1111/pops.12209

Houck, S. C., Conway, L. G., III, Parrow, K., Luce, A., & Salvati, J. M. (2018). An integrative complexity analysis of religious and irreligious thinking. *SAGE Open, 8*(3). https://doi.org/10.1177/2158244018796302

Houck, S. C., Repke, M. A., & Conway, L. G., III. (2017). Understanding what makes terrorist groups' propaganda effective: An integrative complexity analysis of ISIL and Al Qaeda. *Journal of Policing, Intelligence and Counter Terrorism, 12*, 105-118. https://doi.org/10.1080/18335330.2017.1351032

Jordan, K. N., & Pennebaker, J. W. (2017). The exception or the rule: Using words to assess analytic thinking, Donald Trump, and the American presidency. *Translational Issues in Psychological Science, 3*, 312-316. https://doi.org/10.1037/tps0000125

Jordan, K. N., Sterling, J., Pennebaker, J. W., & Boyd, R. L. (2019). Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences of the United States of America, 116*(14), 3476-3481. https://doi.org/10.1073/pnas.1903863116

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin, 129*(3), 339-375. https://doi.org/10.1037/0033-2909.129.3.339

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 8788-8790. https://doi.org/10.1073/pnas.1320040111

McCullough, H. (2020). The diamonds and the dross: A quantitative exploration of integrative complexity in fanfiction. *Psychology of Popular Media Culture, 9*(1), 59-68. https://doi.org/10.1037/ppm0000216

McCullough, H. (2019a). "Hey! Listen!" Video game dialogue, integrative complexity and the perception of quality. *Press Start, 5*, 94-107.

McCullough, H. (2019b). Be complex, be very complex: Evaluating the integrative complexity of main characters in horror films. *Psychology of Popular Media Culture.* Advance online publication. https://doi.org/10.1037/ppm0000266

McCullough, H., & Conway, L. G., III. (2018a). The cognitive complexity of Miss Piggy and Osama Bin Laden: Examining linguistic differences between fiction and reality. *Psychology of Popular Media Culture, 7*, 518-532. https://doi.org/10.1037/ppm0000150

McCullough, H., & Conway, L. G., III. (2019, April). *The integrative complexity of over 200,000 tweets.* Paper presented at the Rensselaer Polytechnic Institute Graduate Research Symposium, Troy, New York, NY, USA.

McCullough, H., & Conway, L. G., III. (2018b). "And the Oscar goes to . . .": Integrative complexity's predictive power in the film industry. *Psychology of Aesthetics, Creativity, and the Arts, 12*(4), 392-398. https://doi.org/10.1037/aca0000149

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015.* Austin, TX, USA: University of Texas at Austin.

Prinsloo, C. F. (2016). *Investigating the influence of individual value systems and risk propensities on decision-making quality in value clashing circumstances* (Doctoral dissertation, University of Pretoria, Pretoria, South Africa). Retrieved from http://hdl.handle.net/2263/61270

Putra, I. E., Erikha, F., Arimbi, R. S., & Rufaedah, A. (2018). Increasing integrative complexity on convicted terrorists in Indonesia. *Social Psychology and Society, 9*(2), 35-45. https://doi.org/10.17759/sps.2018090203

Repke, M. A., Conway, L. G., III, & Houck, S. C. (2018). The strategic manipulation of linguistic complexity: A test of two models of lying. *Journal of Language and Social Psychology, 37*, 74-92. https://doi.org/10.1177/0261927X17706943

Robertson, A., Aiello, L. M., & Quercia, D. (2019, July). The language of dialogue is complex. *Proceedings of the International AAAI Conference on Web and Social Media, 13*, 428-439.

Ross, C. (2019, June 19). What if AI in health care is the next asbestos? *Statnews.com.* Available from https://www.statnews.com/2019/06/19/what-if-ai-in-health-care-is-next-asbestos

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609-612. https://doi.org/10.1016/j.jrp.2013.05.009

Schoonvelde, M., Schumacher, G., & Bakker, B. N. (2019). Friends with text as data benefits: Assessing and extending the use of automated text analysis in political science and political psychology. *Journal of Social and Political Psychology, 7*(1), 124-143. https://doi.org/10.5964/jspp.v7i1.964

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31*, 500-506. https://doi.org/10.3102/0162373709352369

Smith, A. G., Suedfeld, P., Conway, L. G., III, & Winter, D. G. (2008). The language of violence: distinguishing terrorist from nonterrorist groups by thematic content analysis. *Dynamics of Asymmetric Conflict, 1*, 142-163. https://doi.org/10.1080/17467580802590449

Suedfeld, P. (2010). The cognitive processing of politics and politicians: Archival studies of conceptual and integrative complexity. *Journal of Personality, 78*(6), 1669-1702. https://doi.org/10.1111/j.1467-6494.2010.00666.x

Suedfeld, P., & Jhangiani, R. (2009). Cognitive management in an enduring international rivalry: The case of India and Pakistan. *Political Psychology, 30*, 937-951. https://doi.org/10.1111/j.1467-9221.2009.00736.x

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*, 169-178. https://doi.org/10.1037/0022-3514.34.2.169

Suedfeld, P., & Tetlock, P. E. (2014). Integrative complexity at forty: Steps toward resolving the scoring dilemma. *Political Psychology, 35*, 597-601. https://doi.org/10.1111/pops.12206

Suedfeld, P., Tetlock, P. E., & Ramirez, C. (1977). War, peace, and integrative complexity: UN speeches on the Middle East problem, 1947-1976. *The Journal of Conflict Resolution, 21*(3), 427-442. https://doi.org/10.1177/002200277702100303

Thoemmes, F. J., & Conway, L. G., III. (2007). Integrative complexity of 41 U.S. presidents. *Political Psychology, 28*(2), 193-226. https://doi.org/10.1111/j.1467-9221.2007.00562.x

Tetlock, P. E. (1981). Pre- to post election shifts in presidential rhetoric: Impression management or cognitive adjustment? *Journal of Personality and Social Psychology, 41*, 207-212. https://doi.org/10.1037/0022-3514.41.2.207

Tetlock, P. E. (1983). Cognitive style and political ideology. *Journal of Personality and Social Psychology, 45*(1), 118-126. https://doi.org/10.1037/0022-3514.45.1.118

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. https://doi.org/10.1037/0022-3514.46.2.365

Tetlock, P. E. (1985). Integrative complexity of American and Soviet foreign policy rhetoric: A time-series analysis. *Journal of Personality and Social Psychology, 49*(6), 1565-1585. https://doi.org/10.1037/0022-3514.49.6.1565

Tetlock, P. E., Metz, S. E., Scott, S. E., & Suedfeld, P. (2014). Integrative complexity coding raises integratively complex issues. *Political Psychology, 35*(5), 625-634. https://doi.org/10.1111/pops.12207

University of Montana Psychology Department. (2018). *Assessment report*. Retrieved from https://www.umt.edu/provost/faculty/deptreports/CHS/default.php

Vergani, M., & Bliuc, A.-M. (2018). The language of new terrorism: Differences in psychological dimensions of communication in Dabiq and Inspire. *Journal of Language and Social Psychology, 37*(5), 523-540. https://doi.org/10.1177/0261927X17751011

Young, M. D., & Hermann, M. (2014). Increased complexity has its benefits. *Political Psychology, 35*, 635-645. https://doi.org/10.1111/pops.12208