Political Psychology

ISPP
International Society of Political Psychology

Political Psychology, Vol. 35, No. 5, 2014
doi: 10.1111/pops.12021

Automated Integrative Complexity

Lucian Gideon Conway III
The University of Montana

Kathrene R. Conway
The University of Montana

Laura Janelle Gornick
The University of Montana

Shannon C. Houck
The University of Montana

Integrative complexity is a conceptually unique and very popular measurement of the complexity of human thought. We believe, however, that it is currently being underutilized because it takes quite a bit of time to score. More time-efficient computer-based measurements of complexity that are currently available are correlated with integrative complexity at fairly low levels. To help fill in this gap, we developed a novel automated integrative complexity *system designed specifically from the integrative complexity theoretical framework. This new automated IC system achieved an* alpha *of .72 on the standard integrative complexity coding test. In addition, across nine datasets covering over 1,300 paragraphs, this new automated system consistently showed modest relationships with human-scored integrative complexity (average* alpha = *.62; average* r = *.46). Further analyses revealed that this relationship consistently remained significant when controlling for superficial markers of complexity and that the new system accounted for both the differentiation* and *integration components of integrative complexity. Although the overlap between the automated and human-scored systems is only modest (and thus suggests the continued usefulness of human scoring), it nonetheless provides the best automated integrative complexity measurement to date.*

KEY WORDS: complexity, automated

It is becoming increasingly common to take formerly human-scored systems and create efficient, automated systems as proxies. For example, both Walker and Schafer's operational code system and Hermann's Leadership Trait Analysis system have relatively recently produced automated versions (see Hermann, 2005, 2008; Walker, 2003; Walker & Schafer, 2006). The present article describes a similar attempt to automate the human-scored construct *integrative complexity.* Integrative complexity grew out of early work on scoring human samples from Streufert and his colleagues (e.g., Streufert, Suedfeld, & Driver, 1965), but the current archival method used by most researchers was developed and popularized by Peter Suedfeld (see, e.g., Suedfeld, Tetlock, & Streufert, 1992).

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

Why an Automated Version of Integrative Complexity?

Why should we automate this particular construct? We discuss four answers to this question below: (1) it is conceptually unique; (2) it is popular; (3) hand scoring limits its applicability because it is time-consuming; and (4) no validated automated measurement of it is currently available. We now take each of these in turn.

Reason One: Integrative Complexity is a Unique Planet in the Cognitive Complexity Universe

Integrative complexity is only one of many measurements to have "complexity" in the title. So why focus on it particularly? How can it be distinguished from other conceptual approaches to complexity, and why does it matter?

Some semantic clarification is in order. We use the term "cognitive complexity" to denote the larger theoretical construct involving the degree to which people think complexly. Integrative complexity is just one of many hundreds of measurements used over the years to measure this larger construct. Some of those other measurements also have "complexity" in the title; some of those specific measurements are actually called *cognitive complexity*. In the present article, we distinguish the larger theoretical construct from specific measurements with the same name by always attaching the primary developer's name to them, e.g., "*Pennebaker's* cognitive complexity."

At the large construct level, "cognitive complexity" has been ascribed many meanings, but almost all of those meanings have in common the *demonstration of multidimensional (as opposed to a unidimensional) thinking*. Thus, some of the earliest measurements of cognitive complexity involved analyzing the number of independent dimensions that people produced about a given topic using a bipolar rating grid (e.g., Bieri et al., 1966). Other early measurements showed a similar focus on analyzing the dimensional structure of more open-ended statements (Streufert et al., 1965).

So how is integrative complexity distinguished from other methods? To illustrate, we discuss below three dimensions in which IC differs (in varying degrees) from other measurements: (1) IC is deductively generated; (2) it allows for multiple types of complexity; and (3) it scores higher levels of complexity than typical measurements. In doing so, we focus this discussion primarily on IC and two other measurements that are directly relevant to the present article: Hermann's Conceptual Complexity (see Hermann, 2005, 2008) and Pennebaker's Cognitive Complexity (see Pennebaker & King, 1999; Slatcher, Chung, Pennebaker, & Stone, 2007; see also Owens & Wedeking, 2011). Of course, these are not the only three complexity measurements out there, but we discuss them below in more detail because they are popular and well-validated, they are representative of different types of existing measurements used for scoring open-ended materials, and they have (counting the work done in this article on IC) been automated in some form.

Deductive Versus Inductive Approach to Development

Conceptually, there are two very different ways one could design a complexity measurement. First, you could decide up front what the criteria for complexity were (e.g., multidimensional structure versus unidimensional structure), and then you could design a system to measure that directly. This approach, which we loosely refer to as *deductive*, is essentially a direct measurement of a set of predetermined criteria. In the same way someone might say "extroversion means being gregarious" and then sets out to measure *gregariousness*, this approach says (for example) "complexity means multidimensionality" and sets out to measure *multidimensionality*. Both IC and Hermann's Conceptual Complexity were developed in a deductive manner.

An *inductive* approach, by contrast, starts at the other end: It imagines what things complexity ought to predict and then attempts to see what linguistic features do in fact predict those things. For example, Pennebaker's Cognitive Complexity is explicitly inductive in nature. For this measurement, researchers reasoned that cognitive complexity ought to be correlated with class performance and openness to experience (see Pennebaker & King, 1999; Slatcher et al., 2007); they then set out to find linguistic patterns that would be markers of those two things.[1] This approach is thus analogous to saying "extroversion should predict success in social situations" and then seeing what linguistic markers actually *do* predict success in social situations to construct the measurement.

Why this matters. Most measurements of complexity have historically been more deductive than inductive, and so integrative complexity does not stand out substantially in this regard. However, it is worth noting that, although both approaches have strengths and weaknesses, the deductive approach is a *purer* measurement: It is an attempt to directly measure what the construct actually *ought* to be, irrespective of what it may or may not predict in the real world. This deductive approach is partially followed in the present articleby using the same set of criteria that were originally used in designing human-scored IC to also design our automated IC system.

Dialectical Versus Elaborative Formulations of Complexity

A second relevant conceptual distinction is between what Conway et al. (2008) refers to as *dialectical* versus *elaborative* forms of complexity. On the one hand, complexity can be thought of as an attitude of openness to new information. Thus, markers of ambiguity, uncertainty, or a willingness to see multiple perspectives as valid (even if competing) would be considered complexity under this rubric. Conway et al. referred to this kind of complexity as *dialectical* complexity.

On the other hand, multidimensional thinking is not limited to the merely ambiguous or to competing points of view. People can be multidimensional, for example, when defending only one particular perspective about which no ambiguity is felt. Thus, markers of elaboration of a specific viewpoint, multiple dimensions offered without qualification, and several complex arguments in defense of a particular perspective would be considered complexity under this rubric. Conway et al. referred to this kind of complexity as *elaborative* complexity (see Conway et al., 2008, for a more complete discussion).

One of the ways that IC stands out in the cognitive complexity universe is that it explicitly incorporates *both* forms of complexity. Most complexity measurements focus almost exclusively on the dialectical conceptualization of complexity. Hermann's Conceptual Complexity explicitly conceptualizes complexity as the ability to see ambiguity and openness to new information. Similarly, one of the validity markers used in developing Pennebaker's cognitive complexity measurement was its correlation with *Openness to Experience*, which is explicitly a measurement of personality inclinations towards allowing new and/or competing information.

[1] In reality, of course, the distinction between these two approaches is not nearly as sharp as this implies. Many things, for example, predict class performance, and yet one would only call something that predicted class performance "cognitive complexity" if the words that ultimately coalesce to predict that thing seemed relevant to cognitive complexity. On the flip side, the deductive measures would almost certainly not have gained any traction if they did not predict things that it seemed like they ought to predict.

Integrative complexity, in contrast, explicitly incorporates both dialectical *and* elaborative forms of complexity into its rubric. (Though some IC researchers emphasize dialectical forms of IC more than others—see Conway et al., 2008, for a review). As a result of this, IC captures a wider range of complexity forms than the typical measurement, and our automated system developed in this article incorporates multiple markers of both kinds.

Why this matters. This is important because a fairly large percentage of the ways people can be complex are elaborative, and thus any system that leaves out elaborative forms of complexity entirely may miss really important phenomena. For example, consider the finding that psychological extremism is sometimes positively related to complex thinking. Because the complex thinking considered in that case is typically elaborative in nature, systems that did not account for this would completely miss this effect (see Conway et al., 2008). In the present automated instantiation of IC, we explicitly attempt to incorporate both dialectical *and* elaborative markers of complex thinking.

Differentiation Versus Integration

Finally, as previously implied, essentially all measurements of complexity are based in some way on *differentiation*, which means making distinctions among dimensions in the environment. However, some measurements of complexity have added to that the notion of *integration*: At higher forms of complexity people not only differentiate things in the environment, but they also are able to draw connections between those differentiated things.

IC stands out in the cognitive complexity universe as one of the few measurements (and perhaps the only open-ended coding scheme) that explicitly accounts for integration. Conceptual complexity focuses only on differentiation (Hermann, 2005; Suedfeld, Frisch, Hermann, & Mandel, under review). Pennebaker's cognitive complexity, while not explicitly limiting itself to differentiation, nonetheless does not directly account for integration (and all of its word categories are differentiation-focused). This can perhaps be seen in the initial name given to Pennebaker's cognitive complexity measure: "Making distinctions" (see Slatcher et al., 2007).

Why this matters. This distinction is of no small consequence. Conceptually speaking, excluding integration from complexity measurements misses a lot of what cognitive complexity truly *is*. Consider a parable from architecture. Imagine a cul de sac with two buildings side by side, one an old Medieval castle and the other a modern glass skyscraper. That would be more complex than a cul de sac with, say, two identical castles, because different sorts of buildings are more complex than identical buildings. So far, so good. But now imagine that an architectural genius came along and made *one* larger structure out of these two discrepant buildings. This genius built breezeways, columns, arches, and connecting upper floors to seamlessly integrate these two wildly distinct structures into one larger structure. The point of the parable: The larger integrated structure is universally recognized as a more complex structure than the two disparate structures alone. The trusses and archways and such necessary to put two things together are more complex than the things left by themselves. It requires a more complex mind to connect two disparate things than it does to simply recognize the disparate things as disparate. Similarly, complex organisms are not merely more *separate* than simple organisms: In fact, a complex organism is practically defined by the integrative activity of multiple disparate parts. That most complex of organic structures, the human brain, is complex not because it has different independent systems (vision, hearing, smell, language) but precisely because those different independent systems integrate into something like a single stream of thought.

The same thing applies to cognitive structures: Integrating different things is conceptually more complex than simply identifying different things. This *integration* is a hallmark of integrative complexity that is mostly absent from other measurements of complexity and one of the key important advances here: Our automated system, like the human-scored version of IC, explicitly incorporates markers of integration over and above differentiation.[2]

Reason Two: Integrative Complexity is Popular

Although integrative complexity is only one of many possible measurements of cognitive complexity, it is nevertheless the most popular. Its usage spans multiple disciplines, but no single area has been more influenced by integrative complexity than political psychology (see, for example, Conway & Conway, in press; Conway et al., 2012; Conway, Dodds, Hands Towgood, McClure, & Olson, 2011; Conway, Suedfeld, & Clements, 2003; Conway, Suedfeld, & Tetlock, 2001; Conway et al., 2008; Suedfeld, 1985, 1994; Suedfeld & Bluck, 1988; Suedfeld, Conway, & Eichhorn, 2001; Suedfeld, Leighton, & Conway, 2006; Suedfeld & Piedrahita, 1984; Suedfeld & Tetlock, 1976; Suedfeld, Tetlock, & Ramirez, 1977; Suedfeld et al., 1992; Suedfeld, Wallace, & Thachuk, 1993; Tetlock, 1984, 1985, 1986, 1993; Tetlock, Bernzweig, & Gallant, 1985; Thoemmes & Conway, 2007). It has been used by political psychology researchers on topics ranging from international peace (and war), electoral success, political revolutions, international crises, political ideologies, and profiling specific political figures, to name just a few. This large litany of scored materials spans thousands of years and covers multiple continents, nations, and languages (see Suedfeld et al., 2006, for a summary).

The reasons for its wide usage partially overlap with its conceptual uniqueness: Because it is a broad measurement of complexity that captures aspects of complex thinking that some other measurements do not, this makes its potential applicability wider. As a result of this popularity, constructing an automated measure of the construct might also have wide application.

Reason Three: Integrative Complexity Research Is Currently Being Underutilized

However, despite its popularity, we believe the construct has nonetheless been underutilized in recent research. Why? We suggest it is because of a combination of recently widely available computer measurements of other linguistic constructs and the fact that integrative complexity takes quite a bit of time to score by hand (see Suedfeld et al., under review).

To begin with, would-be scorers must take a short training course on scoring integrative complexity, at the end of which they must pass a test set at a very high standard. It often takes many coders to achieve an acceptable level of reliability for a specific project, so to maintain a vibrant lab doing integrative complexity research requires many qualified coders each academic term. Given the turnover among students, this is no small task. Once enough coders are in place for a specific IC project, running that project itself can be time consuming. Before IC coders can begin coding, much time must be spent preparing the materials. Documents for the study must be gathered and divided into codeable paragraphs. From those paragraphs, a random subset from each document must be selected. Then, "blinding" rules must be created and applied to the paragraphs to prevent proper

[2] Historically, other dimensions for characterizing complexity measurements have also been discussed. However, we do not think those dimensions are relevant to our present discussion. For example, one could categorize complexity measurements as focused on process versus outcome. Process measurements are more focused on whether one is actually engaging in an information search, while outcome measurements are focused on whether or not the resulting output was complex in structure. This distinction is immaterial here; all the measurements of interest here are outcome measurements (although Hermann's conceptual complexity measurement is conceptually more inundated with process language, practically speaking it is a measurement of the complexity of the output). Similarly, a distinction has been made between measurements focused on stable personality differences in complexity and those focused on state fluctuations, and in this rubric IC is said to be a state-based measurement and CC a trait measurement. But this distinction is entirely artificial with respect to the measurements themselves, and the degree of state versus trait is in reality an empirical question that has to be settled independently for each construct (and, in fact, research suggests that IC contains a personality component and CC a state component). Thus, it is not relevant to our present discussion.

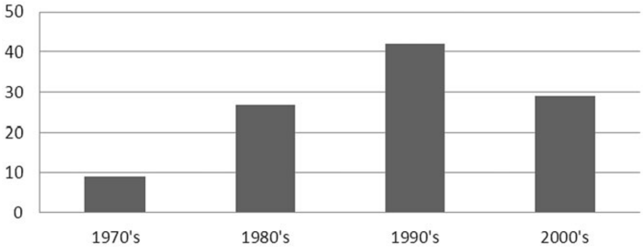

Figure 1. Peer-reviewed journal publications on integrative complexity across the last four decades.

nouns and dates from revealing the author, time period, or context of the document. The actual scoring of paragraphs by coders can be a very time consuming and difficult process in and of itself.

While some of these disadvantages are offset by the fact that one can use almost any available material for research—and as such researchers do not necessarily have to go through the procedure of collecting participant data—nonetheless we believe that the length of time it takes to score materials is, in the modern age of computer measurement, a very real stumbling block for many researchers in using the construct.

To illustrate this potential, we used PsycINFO to search for all peer-reviewed journal articles that contained "integrative complexity" as a keyword. This search yielded over 120 articles. We also mapped these articles in 10-year blocks over the last 40 years. As Figure 1 reveals, the popularity of integrative complexity research appeared to be at its apex in the 1990s. This past decade not only failed to show the continued growth in publications that would be indicative of an expanding area of research, but actually showed a decline in publications to levels roughly the same as the decade prior to the 1990s. While this is still a substantial number of peer-reviewed publications (and is more than are associated with any other specific complexity measurement that we are aware of), we think these data are in fact illustrative of the need for a more efficient measurement system.

Has interest in the construct itself waned? That is of course possible, but we do not think so. We think rather that people have either begun to search for substitutes that are similar but easier to score or just moved on to related areas for which more efficient linguistic measurements do exist. For example, it is worth noting that that in the past decade, other available automated measurements of complexity are often *associated* with integrative complexity (we return in a moment to the question of whether those measurements are *actually* related to human-scored integrative complexity). In a recent conference paper featuring a large discussion of integrative complexity theory and research, Hermann's computer-based Conceptual Complexity measurement was explicitly viewed as a suitable substitute for integrative complexity (Ishiyama, Breuning, & Backstrom, 2010). Sometimes, equating the systems in common parlance is more ambiguous or subtle, as when Dille & Young (2000, p. 588) state that automated conceptual complexity and integrative complexity are "related" or when Slatcher et al. (2007, p. 67) state that Pennebaker's cognitive complexity measure "resembles" integrative complexity. On a different domain, a recent article attempted to provide a pen-and-paper scale for measuring integrative complexity in laboratory participants (Carroll & Bright, 2010). These examples suggest that people are still interested in integrative complexity at some level but are searching for easier ways to incorporate the construct.

In sum, we strongly believe that integrative complexity is a compelling and interesting construct that researchers would use more frequently if it did not take quite so long to score. In a modern academic environment where increased expectation on faculty to publish at greater rates is combined with the availability of multiple automated measurements of many linguistic variables, it is hardly a wonder that integrative complexity research might be currently less utilized than it was before. Thus, we feel an automated version of it might increase the usage of this important construct.

Reason Four: Existing Validated Automated Complexity Measurements Are Not Highly Correlated with Human-Scored Integrative Complexity

Of course, if validated automated systems already existed that could serve as alternatives to integrative complexity, there would be no need for a new system. We are aware of two existing validated automated measurements of cognitive complexity: Hermann's *Conceptual Complexity* (one of seven measurements in her Leadership Trait Analysis system) and Pennebaker's *Cognitive Complexity* (one of multiple measurements in his Linguistic Inquiry and Word Count, or LIWC, system). As mentioned earlier, both of these automated measurements have been discussed in research as being either a direct proxy of, or conceptually related to, integrative complexity (e.g., Dille & Young, 2000; Ishiyama et al., 2010; Slatcher et al., 2007).

However, neither of these computer-based measurements was developed with integrative complexity specifically in mind, and as such, it is hardly surprising that both are correlated at fairly low levels with human-scored IC. Suedfeld et al. (under review) revealed an average correlation between human-scored IC and automated Conceptual Complexity of .21.[3] (And in fact, the correlations between *human-scored* Conceptual Complexity and IC are actually slightly negative, average $r = -.05$, highlighting the fact that the two systems of scoring complexity simply do not overlap that much; see Suedfeld et al., under review). Further, as we demonstrate in this article, the LIWC complexity score is generally correlated at very low levels (average $r = .14$ across nine datasets) with human-scored integrative complexity.

These low correlations are hardly surprising given the conceptual differences between these complexity measurement approaches and integrative complexity (discussed earlier) and the fact that, as a result, neither system was developed to be comparable to integrative complexity.[4] And yet, despite these low correlations, such measurements are sometimes talked about as equivalent to integrative complexity.

Is This Even Possible? Difficulties in Creating Automated Complexity Measurements

So far, we have established that there appear to be good reasons to produce an automated measurement of the integrative complexity construct: It is conceptually unique, popular enough to suggest a potential market for an automated measure, yet time-consuming enough to suggest that researchers might take advantage of the construct in greater numbers if a more efficient automated version were available. And no validated automated measure for it currently exists.

Yet there is often a large gap between the thinking of a thing and the doing of the thing. It would be nice to have an automated integrative complexity measurement, but is it a realistic goal?

If the goal is to obtain near-perfect correlations so that the automated construct can be viewed as a *direct equivalent* to the human-scored construct, then the answer to that question is probably no. As we illustrate below, there are too many inherent difficulties in translating complexity to a computer-based system for that. However, we do believe it is possible to produce a measurement that is a viable *alternative* to integrative complexity—a measurement developed by integrative complexity researchers using an integrative complexity framework and showing modest correlations with human-scored integrative complexity. That is what we believe we have accomplished with our *automated integrative complexity* measurement presented in this article.

[3] Suedfeld et al. (under review) report correlations for three samples. For one of the samples, they do not include an average but rather a range, and we here use the mean of that range to compute this average.

[4] Fairly late in the process of submitting this article, we discovered that the LTA software group had, using Profiler Plus, developed what appears to be a more direct attempt to automate integrative complexity (called "conceptual integrative complexity"). Suedfeld et al. (under review) report a correlation across over 2,500 paragraphs of .27 (only a slight improvement from the .21 we report here for Profiler Plus' automated Conceptual Complexity). We could find no other information on the system; therefore, we do not deal with it directly here.

Why is this so difficult? Any automated system is dependent on word/phrase counts of certain "categories" of words (e.g., more "exclusive" words like "but" equals higher complexity). However, integrative complexity can be achieved in so many ways that it is hard for word counts to capture consistently and adequately complex thinking. Distinctions between word usages that humans see easily are hard to replicate in the template-matching operations that comprise all computer-driven systems. The possible permutations of the context—permutations that often change the complexity-related relevance of the words in question—are almost endless.

In Table 1, we outline some specific obstacles that make mimicking human-scored IC very difficult and some potential solutions. These problems vary in their degree of solvability, which we highlight below.

Problems with General Solutions

Some of these problems have potential solutions that apply across most types of data. Consider, for example, the problem that many connecting words associated with complexity occur frequently in noncomplex usages. For example, the word "but" is used as a marker of complexity in most of the complexity systems currently available. Sometimes "but" *is* a marker of complexity, as in *I like ice cream but I also hate it because it makes my stomach hurt when I eat it.* "But" in this sentence illustrates competing alternative views that are both legitimized by the speaker. On the other hand, consider the statement *Johnny does not like ice cream, but he's totally wrong; ice cream is wonderful.* The "but" in this sentence does not illustrate complexity at all, but rather is a simple rhetorical way of pointing out that their friend is wrong, and they are right. Such black and white thinking is a hallmark of simplicity, not complexity. Human-scored integrative complexity easily makes such distinctions (Suedfeld et al., 1992), but automated systems do not.

While no *perfect* solution to this sort of problem exists, it is nonetheless possible for an automated system to distinguish between these different usages in multiple ways, including (a) only counting the term in question if it appears with other words that indicate it is used as a complexity marker (in this case, "but" would only count in the first instance because it appears in the phrase "but I also"), (b) only counting such words if they appear a lot of times in a short space, and/or (c) ignoring the term (as in the case of "and," which is not correlated with complexity overall).

It turns out that a large percentage of the problems faced when trying to develop an automated complexity system have such generalized solutions. Of course, specifying which exact phrases are typically complex versus those that are not takes a lot of time and requires an extremely large dictionary of complex phrases; but we have developed just such a system. Using these and other methods (please see Table 1 and the methods section), we were able to overcome quite a number of these obstacles.

Problems with Specific Ad Hoc Solutions

Some problems do not lend themselves to easy generalized solutions that apply across most data sets. For example, one of the datasets we scored for automated integrative complexity came from the Nixon/Kennedy presidential debates. As we later demonstrate, this was the only dataset (out of nine) presented here that our automated system did particularly poorly on. Why? Many of the problems had to do with alternative word usages specific to that sociohistorical context. For example, there was a mention of school "integration" which was scored highly by our automated system (because "integration" is usually associated with complexity) but recognized (accurately) as purely descriptive by human coders. Similarly, "balanced" budget, television "debates," and government "resolutions" were all scored as complexity by the automated system in instances where they were obviously descriptive (and thus not complex usages).

Table 1. Common Reasons for Divergence between Automated and Human-Scored Complexity Systems

Reason For Divergence	Human High/ Auto Low	Human Low/ Auto Low	Human Low/ Auto High	Human High/ Auto High	Possible Solutions for Automated System
Automated System Overestimates:					
Different Word Meanings Based on Historical or Study Context	N/A	N/A	"School **Integration**"	"**Integration** of two perspectives"	(1) Probabilistic, and/or (2) Ad hoc word search for specifically problematic instances on a dataset-by-dataset basis
Different Word Meanings Based on Sentence Context	N/A	N/A	"hand me the **other** hammer"	"on the **other** hand"	(1) Probabilistic, and/or (2) only count word as complex when it appears with other contextual markers (e.g., "on the other hand")
Descriptive Versus Conceptual Usages	N/A	N/A	"they went to the store **together**"	"these two ideas taken **together**"	(1) Probabilistic, and/or (2) only count word as complex when it appears with other contextual markers (e.g., "taken together")
Connecting Complexity Words (e.g., "but," "yet") also Frequently Used for Other Purposes	N/A	N/A	"He hates ice cream, **but** he's totally wrong"	"I hate ice cream, **but** I also love it"	Ignore words frequently used for other purposes unless they (1) appear over a certain density threshold or (2) only count them as complex when occurring with other markers (e.g., "but I also" and not "but")
Automated System Overestimates:					
Complex Word Meaning Directly Reversed by Negation Words	N/A	N/A	"this problem is not **complex**"	"this problem is very **complex**"	Exclude common low-complexity negations (e.g., the system does not count "not complex" as complexity)
Some Common Words are Only Low Markers of Complexity No Matter How Much They are Used	N/A	N/A	"**but** he ate the food anyway; **but** he didn't end up liking it"	N/A	Put a cap on the highest complexity score attainable using certain words
Automated System Underestimates:					
Multiple Dimensions Presented with No Linguistic Complexity Marker	"the play was **bitter** and **funny**"	"the play was **bitter** and **irritating**"	N/A	N/A	No easy solution
Simple Word Meaning Directly Reversed by Negation Words	"this problem is not **simple**"	"this problem is **simple**"	N/A	N/A	Include common high-complexity negations (e.g., the system counts "not simple" as complexity)

Note. Human High/Auto Low and Human Low/Auto High are examples of typical divergence in the systems, but are not necessarily indicative of how *Automated Integrative Complexity* ultimately scores the system; probabilistic = method described in the text.

It is difficult to provide a solution to this kind of historical context problem that would apply across all data sets. One cannot simply ignore the word *integration* altogether across all datasets, for example, because in most data sets that would miss legitimate complexity. But in this particular data set, it was not used in a complex way.

Such coarseness is of course to some degree inevitable. But our new system offers one potential solution: It allows researchers to fairly easily make some ad hoc judgments concerning words that might be pervasive problems in a particular data set. Specifically, researchers can peruse the words that were used to score complexity for each paragraph (which are listed by the package), and then enter ad hoc exclusions into the word/phrase dictionary that apply *only* to that dataset. For example, imagine coding a weight-loss intervention for complexity and consistently noting that words like "balance" and "weighing" were counting as complexity, even though they were simply being used descriptively ("when she was weighing herself yesterday . . ."). Researchers can enter in those words as "exclusions" and the automated IC system will then code the entire dataset *as if those words do not indicate complexity*.

Problems with No Easy Solutions

However, some problems appear to have no clear solutions. Sometimes people discuss different dimensions without using any clear linguistic markers of complexity. Consider, for example, that it is very difficult to easily automate the complexity in the phrase "the play was bitter and funny." That phrase is clearly complex because bitter and funny are different dimensions that pull in different directions (and was scored as complex by human scorers in the practice materials used for training here), yet there are no linguistic markers of that complexity.

Indeed, the sociohistorical context issue described above was not the only problem in the Nixon/Kennedy materials. A detailed qualitative analysis of this dataset after the fact suggested that one of the candidates had a tendency to discuss clearly different dimensions in a complex fashion without using any obvious linguistic marker of that complexity. We know of no easy solution for this problem. It is simply one of the costs of using an automated measure of complexity and one of the reasons that researchers should be cautious in using *any* automated measure of this construct.

Reasons for Optimism

Nonetheless, all is not lost. As the main body of our results show, linguistic markers of complexity are generally prevalent enough that an automated system of complexity still makes sense. As one would expect given the inherent obstacles in such an enterprise, our new *automated integrative complexity* system does not show perfect overlap with human-scored integrative complexity. However, it was developed explicitly from integrative complexity principles by integrative complexity researchers, and, as we demonstrate below, we were able to achieve higher correlations with human-scored IC than other automated systems across nine datasets (average *alpha* = .62; average correlation = .46). Indeed, on the "test," human scorers themselves must pass with a .85 *alpha*; an expert scorer, our automated system, while not "passing" the test, achieved a quite respectable *alpha* of .72.[5] While these correlations do not suggest that our automated system is a *direct equivalent* of human-scored integrative complexity, they demonstrate the highest average correlations (to our knowledge) with this complicated and difficult-to-automate construct.

[5] The standard for the test has fluctuated over the years. Sometimes it is been a correlation of .85, sometimes an *alpha* of .85, and sometimes other standards have been applied as well. We here use the *alpha* criterion because it has been the most frequently applied standard over the past 15 years.

Overall Strategy Used to Develop *Automated Integrative Complexity*

In developing the new system, our strategy was part conceptual and part empirical. On the conceptual side, three of the four authors (all experts in integrative complexity research who passed the integrative complexity coding test and have coded for multiple years), using the integrative complexity manual as a guide (Suedfeld et al., 1992), generated large lists of words and phrases that were relevant to complexity. Each word/phrase was scrutinized by the first author, and complete synonym checks were done for each word (and for each resulting synonym), when such a check was feasible.

Further, we used three available datasets to empirically validate the resulting system. These datasets also were used as a feedback loop to generate new words. We describe these datasets as "training" datasets below. Finally, once the system had been completely developed on the three "training" datasets, we further tested the now-developed system on six "untrained" datasets—that is, on datasets that we did not use for training and thus can be considered "new" datasets with respect to the automated system. The centerpiece of this predictive validity approach was the Integrative Complexity Coding test that all new human scorers must pass to become a certified coder of the construct.

We first describe these datasets briefly below. Then we describe the basic idea behind the system that emerged from this conceptual/empirical approach. Finally, we provide tests of the validity of the new system.

Brief Description of Datasets Used in Present Study

Training Datasets

We used three sets to train on, which represented a diverse set of materials that included both student-generated and archival-generated samples.

Integrative Complexity Official Practice Sets (Suedfeld et al., 1992). These 10 sets of paragraphs are available online to those taking the online course in how to code human-scored integrative complexity (see Suedfeld et al., 1992). Each set contains anywhere from 10 to 20 paragraphs (total $N = 156$). These paragraphs are largely generated from archival sources (many but not all of them political in nature) but also include student answers to paragraph completion stems. They are also intentionally constructed to include a wide range of possible complexity scores. As a result, it is an excellent training set, including much diversity in both material type and complexity scores.

Heritability (Conway et al., 2008; Conway et al., 2011). A second training set was a student-based sample that was originally collected for two primary functions (the relationship of complexity to attitude heritability and psychological extremism), each of which resulted in separate publications (Conway et al., 2008; Conway et al., 2011). We here refer to this data set as the Heritability data set (see Conway et al., 2011, for methodological details of the study). The item stems for this study ranged from religious ("organized religion") to political ("death penalty for murder") to social ("being assertive") in nature. As a result of this variability, they provide an excellent range of topics for the current purpose ($N = 310$).

Early Christian Writings. Finally, a third training set was unpublished data on early Christian writings. Writings from the New Testament (first generation) and from subsequent generations of Christian writers were randomly selected (total paragraph $N = 173$).

Untrained Datasets

As with the trained datasets, the untrained sets also represented a diverse set of materials that included both student-generated and archival samples across multiple domains.

Integrative Complexity Official Coding Test (Suedfeld et al., 1992). This is the official coding test for becoming certified for integrative complexity coding ($N = 30$).

Smoking Cessation Study (Conway, Harris, Catley, Conway, & Gornick, in prep.; Harris et al., 2009). Harris et al. (2009) ran a large smoking cessation intervention and transcribed many of the intervention sessions. Conway et al. (in preparation) subsequently coded a subset of paragraphs from the sessions ($N = 240$ paragraphs) for integrative complexity.

2004 Presidential Election Debates between Bush and Kerry (Conway et al., in prep.). Paragraphs were randomly selected from the three 2004 Presidential debates between Bush and Kerry (paragraph $N = 94$).

1968 Presidential Election Debates between Nixon and Kennedy. Paragraphs were randomly selected from the four 1968 Presidential debates between Nixon and Kennedy (paragraph $N = 96$).

2004 Democratic Primaries (Debate 5). Paragraphs were randomly selected from the fifth Democratic Primaries debate from the 2004 election ($N = 75$ paragraphs) for each of the nine participants in the debate.

2008 Presidential Election Campaign between Obama and McCain (Conway et al., 2012). Paragraphs were randomly selected from the 2008 Presidential campaign between Obama and McCain (paragraph $N = 162$), including the three presidential debates, convention speeches, campaign speeches, and a Parade magazine article on patriotism.

Method

The Automated Integrative Complexity System

The Automated IC system produces a score from 1 to 7 using the same scoring rubric and theory as human-scored IC. In both systems, scores of 1 conceptually represent a total lack of differentiation or integration, scores from 2 to 3 represent levels of differentiation, and scores from 4 to 7 represent differentiation *plus* integration.

The Automated IC system accomplishes this task using two fundamental principles: (1) Hierarchical scoring and (2) Probabilistic scoring. We cover each of these briefly here.

Hierarchical scoring. The first principle follows directly from the integrative complexity scoring manual (Suedfeld et al., 1992) because it distinguishes different "categories" of complexity and assigns them different values based on a hierarchical scoring system. At a broad level, the Automated IC system has two sets of word/phrase lists: one set dealing with differentiation, and one set dealing with integration. The system scores in a hierarchical fashion, first scoring differentiation words/phrases (e.g., "on the other hand") and then scoring integration words/phrases (e.g., "in conjunction with"). No matter how many differentiation words are used in a paragraph, the paragraph cannot achieve a score greater than 3 on those words alone. To achieve a score higher than 3, integration words/phrases must be used. This is conceptually identical to human-scored integrative complexity, where only differentiation + integration can receive scores higher than 3 (Suedfeld et al., 1992).

Probabilistic scoring. A second principle deals with probability. Each word or phrase in the dictionary has been scrutinized for the *probability* that it indicates either differentiation or integration. This has been part conceptual (e.g., looking at synonyms and the most common usages of the word/phrase) and part empirical (e.g., empirical analyses of how predictive various phrases/words were of human-scored IC). While this process has taken almost a year, we here summarize only the outcome.

Words/phrases are weighted according to the probability that they would indicate complexity. Some words/phrases are so frequently indicators of complexity, and have few or no low-complexity uses, that even one mention of them deserves full differentiation (e.g., "on the other hand"). Some words/phrases often indicate complexity, but they often also indicate something else that is not complex at all—and these words/phrases get lower scores as a result. The exact score is based on the estimated ratio of complex to noncomplex usages. For example, "apart from" might be used complexly (as in "apart from this reason, there is another reason why . . ."), but it also might be used in a purely descriptive fashion (as in "I do not wish to be apart from you . . ."). As such, one use of "apart from" would receive a score of 2 (even though, by a human scorer, the first example above would receive a 3 and the second a 1). This score represents that there is some probability that the phrase is used complexly but also some probability that it is not.

Producing the final score. The actual score produced on a 1–7 scale is based on these principles. Importantly, this score operates on the same principle as integrative complexity, where the highest score indicated by the various markers is given. Thus, if one "3" word/phrase is in a paragraph and four "2" words/phrases, the paragraph would receive a "3." If, in addition to this, the paragraph has several integration words, it would receive a higher score still, depending on the nature and complexity probability of those integration words.

Controlling for Superficial Markers of Complexity

It has been noted by others (e.g., Suedfeld et al., 1992) that integrative complexity is correlated with some superficial markers of rhetoric (most commonly, paragraph length). Thus, in all analyses comparing the relationship between automated and human-scored complexity, we control for two such markers: Word length and Paragraph length. In particular, we used Pennebaker's measurement of "big words" (which is the percentage of six letter words in the paragraph; see e.g., Pennebaker, Booth, & Francis, 2007) and created a paragraph word count measure. As seen below, our system is clearly capturing quite a bit of variance above and beyond these superficial markers (and indeed, the markers rarely change the strength of the effect much at all).

Alternative Automated Complexity Measurement: LIWC

For comparison, we also ran the same exact analyses for another automated measurement of the complexity of rhetoric: that derived from Pennebaker's LIWC system (see, e.g., Newman, Pennebaker, Berry, & Richards, 2003; Pennebaker & King, 1999; Pennebaker, 2011; Slatcher et al., 2007). This score is the aggregate of several standardized scores, including the percentage of "exclusive" words, the percentage of "negation" words, and the percentage of "inclusive" words (inverse scored).

In doing so, we are not attempting to set up a head-to-head comparison of the *usefulness* of each system. Indeed, as we have emphasized, the LIWC complexity measurement was not constructed as an attempt to produce an integrative complexity measurement at all, and there are multiple valid ways to approach measuring rhetoric complexity. Given our goal of specifically producing an automated system from within an integrative complexity framework, it is hardly surprising that our measurement performs better at such a task.

Our reasons for including an alternative system are different: We merely attempt to show that our measurement does predict human-scored IC better than another complexity system that is currently available and widely used. (We rely on Suedfeld et al., under review, for comparison figures with Conceptual Complexity; Suedfeld et al. report correlations that average to .21 for that system). As we amply demonstrate, our automated IC system does indeed far exceed this popular, valid, and currently available complexity measure in its ability to predict human-scored IC. This demonstration does not, could not, and is not intended to invalidate the LIWC complexity system, any more than prior demonstrations of low correlations between IC and Hermann's automated Conceptual Complexity measure (Suedfeld et al., under review) invalidate that storied measurement; rather, it is merely intended to show that our system has *unique usefulness* as a measurement of automated

Table 2. Primary Results: Relationship Between *Automated Integrative Complexity* and *Human-Scored Integrative Complexity* Across Trained and Untrained Datasets

	N	Correlation	Control: Word #	Control: Word Size	*alpha*
Training (Total)	**639**	**.56**	**.41**	**.56**	**.72**
Practice Sets	*156*	*.61****	*.51****	*.61****	*.76*
Heritability	*310*	*.49****	*.29****	*.48****	*.65*
Christian Writings	*173*	*.59****	*.44****	*.58****	*.74*
Untrained (Total)	**697**	**.41**	**.30**	**.41**	**.58**
IC Coding Test	*30*	*.57****	*.51***	*.57****	*.72*
Smoking	*240*	*.50****	*.30****	*.50***	*.66*
Nixon/Kennedy	*96*	*.18^*	*.13*	*.18^*	*.30*
2004 Primaries	*75*	*.42****	*.34***	*.42****	*.55*
Obama/McCain	*162*	*.46****	*.24***	*.46****	*.63*
Bush/Kerry	*94*	*.34****	*.29***	*.34***	*.54*
All Datasets Total	**1336**	**.46**	**.34**	**.46**	**.62**

***$p < .001$; **$p < .01$; *$p < .05$; ^$p < .15$

integrative complexity for those researchers (like us) who are primarily interested in that approach to rhetoric complexity.

Primary Results: Correlations with Human Scorers

Our key analyses focus on the relationship between automated and human-scored integrative complexity within each dataset, using the paragraph as the unit of analysis. Because past researchers have sometimes used different metrics for reliability, for ease of cross-study comparison, we present both correlations and *alphas* for all results.

The full results are presented in Table 2. As can be seen there, automated integrative complexity was consistently modestly related to human-scored IC. Overall, the average *alpha* across datasets was .62, and the average correlation was .46. This correlation typically did not change too much (with every significant correlation remaining significant) when controlling for the two superficial complexity markers (word count, word length): Overall, the average .46 correlation dropped to .34 when controlling for word count and remained at .46 when controlling for word length. It is clear from Table 2 that meaningful variance was accounted for by the automated IC system that was above and beyond superficial markers of complexity.

As one would expect, the automated system did somewhat better on the training materials than the untrained materials. However, the gap was not overly large. This is important because the purest test of the automated system's predictive validity is on the six datasets which were not used for training (naturally, when one designs a system to do well on three sets of data, there will of course be lower scores on new data). Indeed, the overall average *alpha* for the six untrained datasets was .58, and the overall correlation was .41. (Please see Table 2.) These analyses suggest a modest-yet-consistent overlap between automated IC and human-scored IC.

The performance of the system on the Coding Test is particularly noteworthy. This is the test that all coders must pass in order to become a "certified" human scorer. While our automated IC system did not "pass" the test, it achieved a quite respectable *alpha* of .72 (and it is worth remembering here that we did not use the test for "training," so this is a measurement of its true predictive validity).

As can be seen in Table 3, automated integrative complexity was superior to the LIWC complexity measurement. The LIWC complexity measurement had an average *alpha* across datasets of .23 and an average correlation of .14. By every metric measured here, the LIWC's relationship with

Table 3. Relationship Between *LIWC's Complexity* Measure and *Human-Scored Integrative Complexity* across Trained and Untrained Datasets

	N	Correlation	Control: Word #	Control: Word Size	*alpha*
Training (Total)	**639**	**.17**	**.15**	**.18**	**.28**
Practice Sets	*156*	*.05*	*.10*	*.05*	*.09*
Heritability	*310*	*.15***	*.14**	*.15***	*.26*
Christian Writings	*173*	*.32****	*.22***	*.33****	*.49*
Untrained (Total)	**697**	**.12**	**.03**	**.12**	**.20**
IC Coding Test	*30*	*.19*	*–.04*	*.19*	*.32*
Smoking	*240*	*.19***	*.12^*	*.19***	*.32*
Nixon/Kennedy	*96*	*.02*	*.01*	*.01*	*.04*
2004 Primaries	*75*	*.05*	*–.19*	*.09*	*.10*
Obama/McCain	*162*	*.14^*	*.16**	*.15^*	*.25*
Bush/Kerry	*94*	*.10*	*.09*	*.09*	*.18*
All Datasets Total	**1336**	**.14**	**.07**	**.14**	**.23**

***$p < .001$; **$p < .01$; *$p < .05$; ^$p < .15$

human-scored IC was substantially lower than our automated IC measure. Please see Tables 2 and 3 for complete results.[6]

Differentiation Versus Integration

One of the advantages of the automated system is the ease of separating differentiation from integration. Indeed, the automated IC system automatically produces separate scores for words/phrases related to differentiation and those related to integration.

For our purposes, it is worth using this capability to ask the question: Do both parts independently contribute to the predictive abilities of the automated IC score? As discussed earlier, one of the ways human-scored integrative complexity is conceptually unique is its incorporation of integration over and above differentiation, and we designed the automated IC system to mimic this conceptual uniqueness. Thus, it is worth finding out if the integration component is contributing anything to prediction.

It is worth noting that we would not expect integration to be as correlated with human scoring as we would differentiation, for at least two related reasons: (1) Fewer paragraphs have integration. One cannot predict variation that does not exist: To the degree that integration is relatively rare, correlations will artifactually go down as a result of this lack of variability. (2) Differentiation is logically necessary for integration. Thus, one would theoretically expect that differentiation would be involved in all complex paragraphs, but integration in only a subset. A logical consequence of this is that differentiation ought to contribute more to the overall score than integration.

[6] A different way of analyzing the reliability of the automated system would be to compare across datasets using the dataset as the unit of analysis. Such an analysis would tell us, writ large, whether or not datasets which generally contained more complexity were scored as such by the automated system (compared to whole datasets with less complexity). Although perhaps less trustworthy in some regards than within-dataset analyses (in part because across-dataset comparisons have coder confounds due to different coders scoring each set), it is still noteworthy to point out that the overall correlation across datasets between human-scored and Automated IC is .82, $p < .01$ (*alpha* = .90). This number is, as one would expect, higher for the trained ($r = .97$, *alpha* = .99) than the untrained ($r = .70$, *alpha* = .83) datasets (and neither of the within-type analyses gained statistical significance, hardly surprising with an n of 6 and 3, respectively), but overall this corroborates the story presented here: From whatever angle of approach one uses, our automated system is generally modestly to-strongly related to human-scored IC. This is further corroborated by the close approximation revealed in terms of absolute scores, with two datasets showing nearly identical absolute scores (2.03 versus 2.02, 1.93 versus 1.90) and many of the other scores being very close (average absolute difference = 0.25).

Table 4. Standardized Beta for Automated Differentiation and Automated Integration Entered as Simultaneous Predictors for Human-Scored Integrative Complexity

	Differentiation	Integration
Training Sets (average):	***.48***	***.23***
Practice Sets	*.37****	*.40****
Heritability	*.50****	*.08∧*
Early Christian Writings	*.57****	*.20****
Untrained Sets (average):	***.34***	***.22***
Coding Test	*.50***	*.38**
Smoking Cessation	*.46****	*.15***
Nixon/Kennedy	*.22**	*–.01*
Democratic Primaries	*.21**	*.34****
Obama/McCain	*.33****	*.33****
Bush/Kerry	*.31***	*.13*

Note. ∧*p* < .15; **p* < .05; ***p* < .01

To test the additive value of differentiation and integration, we used a very conservative approach: We ran regressions for each dataset, simultaneously entering both differentiation and integration as predictors of human-scored IC. Results are presented in Table 4. As can be seen there, for both trained and untrained materials, differentiation was a stronger predictor of human-scored IC than integration (as expected). However, it is clear that for most datasets, integration is contributing additional variance above and beyond differentiation. The average standardized *beta* for trained datasets was .48 for differentiation and .23 for integration; for untrained datasets, differentiation was at .34 and integration at .22.

Across all datasets, differentiation contributed significant additive predictive power in all nine cases, while integration was significant in all nine cases (and one other approaching significance). It is worth remembering that this is a very conservative test because it removes any variance shared between differentiation and integration.

As expected, integration showed more variability across sets. This is directly attributable to the likelihood of a dataset having integration at all, approximated here by two different methods: the highest human-coded score for the dataset (higher scores indicating a greater baseline for possible integration) and the percentage of paragraphs that were scored by humans as containing integration (as indicated by an average among coders > 3). As a rule, for sets with more integration on human scoring, our automated integration measurement showed stronger predictive validity overall. This is indicated by macrolevel correlations (using the dataset as the unit of analysis) between (1) each indicator of integration likelihood and (2) the strength of the relation between automated integration and human-scored IC: .75 for the *highest score* and .68 for the percentage-over-3 measure. Thus, as the likelihood of integration goes up, the additive predictive value of the integration component of the IC measure goes up as well.

Overall, these data clearly suggest the importance of *both* parts of the automated IC formula. Indeed, for sets that intentionally included paragraphs with integration in them (the practice sets and the IC coding test), the integration component was either stronger or close to as strong as the differentiation component (.38 and .40; see Table 4). Thus, although as with any data variability existing across datasets, the overall picture is clearly supportive of the effectiveness and necessity of both parts.

Overlap of Patterns Between Human-Scored and Automated IC

So far, we have demonstrated that, across multiple types of datasets, (1) our automated IC measure is consistently but modestly related to human-scored IC, (2) this relation is not accounted for by superficial markers commonly assumed to be related to IC, (3) both the integration and differentiation components are important in its predictive validity, and (4) this relation is much stronger than another automated complexity measurement from a commonly used automated system.

We also looked to see what would happen if one used the automated system to try and replicate findings from the datasets used here that produced meaningful results with the human-scored measurement of integrative complexity. We summarize these results in Table 5.

We organize the table by the success of the attempt in replicating a human-scored finding. A "successful replication" means the data more or less completely replicated the pattern of the human-scored measurement, both in terms of pattern and in terms of inferential significance. A "partially successful replication" means that the pattern was mostly the same but showed some differences, and/or the pattern was in the same direction but inferentially weaker. An "unsuccessful replication" means that the automated measurement did not replicate the pattern at all.

This is, in many ways (unlike testing the direct correlation between automated and human-scored systems), a somewhat indirect approach that is more of a rhetorical exercise. Naturally, when two things are correlated at an average of .46 (which was the average correlation across all trained and untrained datasets between automated and human-scored IC), they will of course sometimes show the same pattern of results and sometimes not.

Thus, what one would reasonably expect from such a comparison, given measurement error, is that the automated system would show some reasonable overlap in terms of its predictive validity. As we see in Table 5, that is exactly what we find.[7] Two of the replication attempts were essentially identical replications, two of them did not replicate the pattern at all, and the rest showed varying degrees of success in partially replicating the findings.

Discussion

We created a new automated system from within an integrative complexity framework, and the resulting *automated IC* measure was consistently but modestly correlated with human-scored IC across nine datasets containing over 1,300 paragraphs. These datasets were very diverse in content, length, and average complexity score. Further, our analyses suggest that the predictive power of automated IC remains largely unaltered when accounting for superficial markers of paragraph structure (paragraph length and word size) and for both differentiation and integration, and it is much stronger than an already existing automated complexity measure (and stronger than prior reports of a second existing measure; Suedfeld et al., under review).

[7] We also performed other tests as well across datasets that we are not reporting here. Sometimes those tests did not yield a significant effect for either human-scored or automated IC; sometimes they yielded a significant effect for both; sometimes the effect itself, while occurring in some manner for one or the other measure, did not seem trustworthy. The spirit of those additional analyses is that sometimes the two measures yield the same results, and sometimes they do not; the likelihood of them yielding the same analyses seems to go up as the strength and/or theoretical soundness of the effect goes up. As this spirit is captured here, we felt it would unnecessarily bog the article down to engage in too much detail. We acknowledge, however, that we are naturally more interested in evidence *for* the validity of our automated system than evidence against it. While we are not intentionally hiding or skewing our results, it is probably true that our efforts and methods somewhat subjectively overestimate the real likelihood of the two systems yielding an identical pattern. However, this possible overestimation does not apply to the actual raw relationship between the two systems (e.g., *alphas* and correlations), as in that case, we have reported all the data analyses we have done.

Table 5. Summary of Replication Attempts

Dataset	Outcome	Human-Scored Outcome	Automated Outcome
Early Christian Writings	Successful Replication Attempt	Doctrinal/Apologetic statements significantly greater than narrative statements, even when controlling for word count	Doctrinal/Apologetic statements significantly greater than narrative statements, even when controlling for word count
Smoking Cessation	Successful Replication Attempt	Significant difference in complexity for smoking cessation outcomes between Successful Attempters and Unsuccessful Attempters, even when controlling for word count	Significant difference in complexity for smoking cessation outcomes between Successful Attempters and Unsuccessful Attempters, even when controlling for word count
Bush/Kerry Debates	Partially Successful Replication Attempt	Significant Topic × Person interaction	Significant Topic × Person interaction; some similarities in the pattern compared to human-scoring, but some differences as well
Obama/McCain Debates (Mean Differences Between Candidates)	Partially Successful Replication Attempt	Obama significantly more complex, but this was almost entirely due to complexity on domestic (and not foreign) topics	Pattern very similar to human-scored measurements, but inferentially weaker
Meta-Analyses of three Presidential Debates	Partially Successful Replication Attempt	Obama and Nixon near the top of six candidates in complexity; McCain and Bush near the bottom	Candidate-Level (n = 6) correlation between automated and human systems = .41 (alpha = .58); Obama and Nixon near the top of six candidates in complexity; McCain and Bush near the bottom; main divergence was Kennedy
Democratic Primaries	Partially Successful Replication Attempt	Braun and Kucinich near the top ten candidates in complexity; Lieberman, Kerry, and Gephart near the bottom	Candidate-Level (n = 10) correlation between automated and human systems = .48 (alpha = .65); Braun and Kucinich near the top of ten candidates in complexity; Lieberman, Kerry, and Gephart near the bottom; main divergences were Clark and Sharpton
Heritability	Unsuccessful Replication Attempt	Heritability significantly positively correlated with complexity	Heritability non-significantly negatively correlated with complexity
Obama/McCain Debates (Predictive Validity of Voter Preference)	Unsuccessful Replication Attempt	Complexity on foreign topics significantly increased the likelihood people would vote for McCain	Effect non-significant and near-zero

Note. The Nixon/Kennedy Dataset did not yield any interpretable significant effects for either the human-scored or automated measurement, and therefore no "replication" can be meaningfully discussed (except in the context of the meta-analysis of presidents).

The Automated IC System as Integrative Complexity?

Can this automated IC be used as a *direct equivalent* to human-scored integrative complexity? The answer to this is obviously "no." Our average correlation for untrained data was .41, meaning that, although the overlap is nontrivial, a large percentage of the variance is still unaccounted for.

So, given this, why call it automated integrative complexity and consider it a part of that rubric? And why should researchers use this system? Let's take each of these questions in turn.

Why consider this an IC measurement? This automated measurement was designed solely around the integrative complexity construct by researchers trained in the IC tradition and who focus on IC research. The IC manual (Suedfeld et al., 1992) was used as the primary source in its construction; the theory that underlies human-scored IC also served as the bedrock underlying the automated measurement. Although the correlations suggest they are not the exact same measurement, they are conceptually designed around the exact same set of criteria. The automated measurement is, in a sense, a *coarser* measurement of the same construct. Yet the fact that *conceptually* they are attempts to measure the same construct is undeniable.

So we consider automated IC a measurement of integrative complexity for the same reason that researchers, at a larger level, have often used the term "complexity" to describe even uncorrelated measurements (and certainly measurements that are less correlated than ours is with IC): because they share a clear conceptual overlap (see, e.g., Conway et al., 2001). After all, no one would argue that Hermann's conceptual complexity or Pennebaker's LIWC complexity measurements are not *really* complexity measurements, even though such complexity measurements are not always highly related to each other. We argue the same thing for automated IC at a more specific level: Not only is it conceptually a complexity measurement, it is conceptually an *integrative complexity* measurement. We feel that, overall, given the difficulty of achieving correlations consistently with IC, and the direct conceptual overlap, combined with our modest correlations (average = .46), it is justified to put this construct under that rubric.

Why use automated IC? It is also worth placing this discussion in the context of an issue we raised in the introduction: the potential need for automated integrative complexity measurements. Given that IC is the most widely used measurement of rhetoric complexity, and given the increasing trend to use computer measurements of complexity regardless, it makes sense to have an automated measurement of IC that is as *comparable as possible*. It seems to us that our measurement, while not perfect, may be the best on the market right now for this specific purpose. Thus, if researchers are going to move forward using automated measurements that they wish to be as equivalent to IC as possible, we think our measurement fits the bill.

Further, it is important to remember that the disadvantages of the automated IC system are in part offset by one glaring advantage: namely, that it can score a *lot* more material than the human-scored system. In our validity tests, we were more likely to replicate an effect that was larger, more obviously theoretically related to complexity, and/or overwhelmingly statistically significant. Remember that our data were all originally set up for human scoring and thus represent a relatively small capacity of what an automated system would be able to accomplish. As a result, the *coarseness* of measurement inherent in translation to an automated system could in part be offset by the larger *N* it would produce (rather easily) in an effort to find a real effect. (It is also worth noting that almost all human-scored projects take a random sample of that material; whereas, using our automated IC system, no sampling is necessary—*all* the available data can be scored.) It is the classic quantity-versus-precision trade-off. Given the amazing amount of electronic material currently available to researchers in an ever-expanding electronic universe, this advantage is (we deem) far from trivial.

The automated measurement also provides some additional value above and beyond human-scored IC. For example, because the automated IC package gives scores for differentiation and integration separately, this affords researchers a powerful and simple tool: the ability to easily test hypotheses that separate differentiation from integration. Prior to this, such a separation has been difficult and cumbersome and not very clean. Automated IC allows for a very clean and easy breakdown of the degree a paragraph contained differentiation words versus integration words. For example, consider the relationship between affiliation and integrative complexity (see Thoemmes & Conway, 2007). It is possible that this relationship has more to do with integration (a sense of connecting things) than with differentiation, and the automated IC system allows for a quick and easy test of this hypothesis.

Using the Human Mind . . . and Large Samples

Given the inherent difficulties with converting such a complex human-scored system to an automated one, we think it vital that researchers use this system judiciously. Some coarseness in any measurement (human or automated) is to some degree inevitable. But researchers can take steps to reduce the scientific consequences of it. First and foremost, researchers should use *large, diverse samples* of materials for scoring. To name just one specific advantage, using larger samples greatly reduces the likelihood that a few "descriptive" instances of a typically complex word will skew the results. Consider the problem that we discussed in the introduction of "school integration" being incorrectly scored as complexity in the Nixon/Kennedy dataset. Because that dataset had only 96 paragraphs, only a few instances of such mis-scored phrases can skew the dataset. However, as we have suggested, such small sampling is a poor use of an automated system; and if instead of 96 paragraphs, 6,000 paragraphs (for example) are scored across multiple contexts, this decreases the odds that a few instances of one particular mis-scored phrase will matter to the overall results.

This is also where the Automated IC package's allowance for researchers to create ad hoc exclusions comes in handy. By scanning the list of words counted as complex, researchers can focus their attention only on those words that seem to make up a relatively large *proportion* of complexity in the data set. In this way, researchers can be relatively efficient in finding potential problems that are unique to the sociohistoric or research context.

It is worth noting that, in any system across any context, there is a possibility for error. We have demonstrated here that, used as is, our system has a reasonable likelihood of at least coarsely measuring something that can meaningfully be considered under the integrative complexity rubric. Nonetheless, we think some small (and not very time consuming) safeguards by researchers will help minimize the scientific difficulties inherent in an automated system. In other words, even when using an automated system, there is no substitute for the human mind.

Concluding Thoughts

Integrative complexity, despite its popularity, may be underutilized because it is time-consuming to score. We believe our *automated integrative complexity* system is a scientifically valid alternative to human-scored integrative complexity that is far more efficient. We use *alternative*, and not *equivalent*, purposefully. In our estimate, nothing can *replace* human-scored IC directly, and we ourselves still plan to use human scorers for smaller projects. However, when large amounts of electronic data are available, we believe our automated system provides a valid measurement of IC that was developed from the same principles and shows modest correlations with the human measurement.

ACKNOWLEDGMENTS

The authors are especially grateful to Peter Suedfeld for his insightful comments on a prior draft of this article and for his help in developing this system. Further, we would like to thank the following people for help with data collection, coding, and/or entry: Kristy Schneider, Amy Allison, Andrea Wogamon, Stacey McClure, Kirsten Hands Towgood, Daniel Dodds, May Green, Emmy Erickson, Emma Petroff, Heather Hollinder, Chloe Ross, Treva Bittinger, Jessica Dallman, Mary Hammes, April Webber, Mikyla Veis, Roger Johnson, and Katie Davey. The project described was partially supported by Award Number R01CA107191 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. The funder had no role in the conduct of the study or preparation of this article. The automated system discussed in this article can be downloaded at http://psychweb.psy.umt.edu/conway. Correspondence concerning this article should be sent to Lucian Gideon Conway, III, Psychology Department, University of Montana, Missoula, MT 59812. E-mail: luke.conway@umontana.edu.

REFERENCES

Bieri, J., Atkins, A. L., Briar, S., Leaman, R. L., Miller, H., & Tripodi, T. (1966). *Clinical and social judgement: The discrimination of behavioral information*. New York: Wiley.

Carroll, J., & Bright, A. (2010). Integrative complexity of public beliefs toward wildfire management: Development of a scale. *Journal of Applied Social Psychology, 40*, 344–359.

Conway, L. G., III, & Conway, K. R. (in press). Cognitive complexity. In D. Christie (Ed.), *Encyclopedia of peace psychology*. Hoboken, New Jersey: Wiley-Blackwell.

Conway, L. G., III, Dodds, D., Hands Towgood, K., McClure, S, & Olson, J. (2011). The biological roots of complex thinking: Are heritable attitudes more complex? *Journal of Personality, 79*, 101–134.

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.

Conway, L. G., III, Harris, K. J., Catley, D., Gornick, L. J., & Conway, K. R. (manuscript under review). Predicting smoking cessation from client complexity during treatment.

Conway, L. G., III, Suedfeld, P., & Clements, S. M. (2003). Beyond the American reaction: Integrative complexity of Middle Eastern leaders during the 9/11 crisis. *Psicologia Politica, 27*, 93–103.

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 (Ed.), *Peace, conflict, and violence: Peace psychology for the 21st century* (pp. 66–75). Upper Saddle River, NJ: Prentice Hall/Pearson Education.

Conway, L. G., III, Thoemmes, F., Allison, A. M., Towgood, K. H., Wagner, M. J., Davey, K., et al. (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.

Dille, B., & Young, M. D. (2000). The conceptual complexity of presidents Carter and Clinton: An automated content analysis of temporal stability and source bias. *Political Psychology, 21*, 587–596.

Harris, K. J., Golbeck, A. L., Cronk, N., Catley, D., Conway, K. R., & Williams, K. (2009). Timeline follow-back versus global self-reports of tobacco smoking: A comparison of findings with infrequent smokers. *Psychology of Addictive Behaviors, 23*, 368–372.

Hermann, M. G. (2005). Assessing leadership style: A trait analysis. In *The Psychological assessment of political leaders*, edited by Jerrold Post. Ann Arbor: University of Michigan Press.

Hermann, M. G. (2008). Using content analysis to study public figures. In A. Klotz & D. Prakash (Eds.), *Qualitative analysis in international relations*. New York: Palgrave.

Ishiyama, J. T., Breuning, M., & Backstrom, J. (2010). *Talking it over: Parliamentary debates, conceptual complexity and post conflict Kenya*. Paper presented at the annual meeting of the International Studies Association, New Orleans LA.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J.M. (2003). Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin, 29*, 665–675.

Owens, R. J., & Wedeking, J. P. (2011). Justices and legal clarity: Analyzing the complexity of Supreme Court opinions. *Law & Society Review, 45*(4), 1027–1061.

Pennebaker, J. W. (2011). Using computer analyses to identify language style and aggressive intent: The secret life of function words. *Dynamics of Asymmetric Conflict, 4*, 92–102.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count (LIWC2007)*. Austin, TX: www.liwc.net.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*, 1296–1312.

Slatcher, R. B., Chung, C. K., Pennebaker, J. W., & Stone, L. D. (2007). Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality, 41*(1), 63–75.

Streufert, S., Suedfeld, P., & Driver, M. J. (1965). Conceptual structure, information search, and information utilization. *Journal of Personality and Social Psychology, 2*(5), 736–740.

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. (1994). President Clinton's policy dilemmas: A cognitive analysis. *Political Psychology, 15*(2), 337–349.

Suedfeld, P. (2010). The cognitive processing of politics and politicians: Archival studies of conceptual and integrative complexity. *Journal of Personality, 78*, 1669–1702.

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

Suedfeld, P., Conway, L. G., III, & Eichhorn, D. (2001). Studying Canadian leaders at a distance. In O. Feldman & L. O. Valenty (Eds.), *Profiling political leaders* (pp. 3–19). Westport, CT: Praeger.

Suedfeld, P., Frisch, S., Hermann, M., & Mandel, D. (under review). The complexity construct in political psychology: Personological and cognitive approaches. *Manuscript under review*.

Suedfeld, P., Leighton, D. C., & Conway, L. G., III. (2006). Integrative complexity and cognitive management in international confrontations: Research and potential applications. In M. Fitzduff & C. Stout (Eds.), *The psychology of resolving global conflicts: From war to peace: Nature vs. nurture* (Vol. 1, pp. 211–237). Westport, CT: Praeger Security International.

Suedfeld, P., & Piedrahita, L. E. (1984). Intimations of mortality: Integrative simplification as a precursor of death. *Journal of Personality and Social Psychology, 47*(4), 848–852.

Suedfeld, P., & Tetlock, P. (1976). 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. *Journal of Conflict Resolution, 21*(3), 1–16.

Suedfeld, P., Tetlock, P. E., & Streufert, S. (1992). Conceptual/integrative complexity. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 393–400). New York: Cambridge University Press.

Suedfeld, P., Wallace, M. D., & Thachuk, K. L. (1993). Changes in integrative complexity among Middle East leaders during the Persian Gulf crisis. *Journal of Social Issues, 49*(4), 183–199.

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. (1985). Integrative complexity of American and Soviet foreign policy rhetoric: A time-series analysis. *Journal of Personality and Social Psychology, 49*(6), 1565–1585.

Tetlock, P. E. (1986). A value pluralism model of ideological reasoning. *Journal of Personality and Social Psychology, 50*(4), 819–827.

Tetlock, P. E. (1993). Cognitive structural analysis of political rhetoric: Methodological and theoretical issues. In S. Iyengar & W. J. McGuire (Eds.), *Explorations in Political Psychology* (pp. 380–405). Durham, NC: Duke University Press.

Tetlock, P. E., Bernzweig, J., & Gallant, J. L. (1985). Supreme Court decision making: Cognitive style as a predictor of ideological consistency of voting. *Journal of Personality and Social Psychology, 48*(5), 1227–1239.

Thoemmes, F. J., & Conway, L. G., III (2007). Integrative Complexity of 41 U.S. Presidents. *Political Psychology, 28*(2), 193–226.

Walker, S. (2003). Operational code analysis as a scientific research program. In C. Elman &M. Elman (Eds.), *Progress in international relations theory* (pp. 245–276). Cambridge, MA: MIT Press.

Walker, S., & Schafer, M. (2006). Theodore Roosevelt and Woodrow Wilson and cultural icons on U.S. foreign policy. *Political Psychology, 28*, 747–776.