Political Psychology

Political Psychology, Vol. 35, No. 5, 2014
doi: 10.1111/pops.12209

Automated Integrative Complexity: Current Challenges and Future Directions

Shannon C. Houck
University of Montana

Lucian Gideon Conway III
University of Montana

Laura Janelle Gornick
University of Montana

Automating integrative complexity is fraught with many challenges. To address these challenges, we discuss the tension between a specificity approach and a more flexible multiple-pass approach, the multifaceted nature of the complexity construct, the gold standard for complexity measurement, the difficulty of human scoring and its consequences for automation, and some ways forward for creating the best complexity measurements. In so doing, we present new data demonstrating (1) initial evidence for the validity of a new automated system for measuring two different forms of complexity (elaborative and dialectical), (2) the danger of constructing measurements in a purely ad hoc fashion that ignores prospective testing, (3) human-to-computer correspondence is in part a function of human-to-human correspondence, (4) human-to-computer correspondence increases systematically as one uses tests with larger units of analysis, and (5) the lack of correspondence of different systems (both human and automated) may occur in part because they were designed for different units of analysis.

KEY WORDS: integrative complexity, automation, computer scoring

Automating integrative complexity is fraught with many challenges. We would first like to express our sincere gratitude to Suedfeld and Tetlock (2014), Tetlock, Metz, Scott, and Suedfeld (2014), and Young and Hermann (2014) for tackling some of these challenges. Their contributions in this symposium spur an important dialogue about the current state of automating complexity and its future possibilities. These collaborative efforts will help fine-tune the automation of complexity.

In this concluding article, we address some of the issues raised in this symposium about how best to overcome the difficulties inherent in automating complexity scoring. Specifically, we address (1) the flexible multiple-pass approach elaborated on by Young and Hermann (2014) as compared to our own approach, (2) complexity as a multifaceted construct, (3) the gold standard for complexity measurement, (4) the difficulty of human scoring and its consequences for automation, and (5) some ways forward for all of us as we strive to create the best complexity measurements. Along the way, we present some additional data that bear on these issues.

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

Specificity Versus Flexibility in Automation

At a broad level, this symposium has highlighted that there are at least two different approaches for automating integrative complexity. Our approach focuses on the specificity of the words via a template-matching system. This approach scores exact words or phrases that are then probabilistically associated with specific complexity levels. In contrast, the approach discussed by Young and Hermann (2014) relies more on flexibility. It has rules that incorporate a broader range of possible word groupings. This flexibility approach applies a variety of algorithms that determine the likelihood that language is complex, based in part on the positioning of typical complexity indicators in relation to other words or indicators.

There is an inherent tension in the strengths and weaknesses of these two approaches. Specificity approaches may miss certain complex phrases captured by flexibility approaches, but they also allow for a greater confidence in the words/phrases that are scored. Flexibility approaches may capture complexity missed by a template-matching approach, but they also create additional noise leading to incorrect scoring precisely because of the flexible rule. Borrowing from Young and Hermann's set of examples, we illustrate this tension below.

Sequence Interruptions in Specific and Flexible Systems

Consider the issue of sequence interruptions raised by Young and Hermann. As they correctly state, our specificity approach is limited to discrete words/phrases, and this leads to systematic misses in texts with sequence interrupters. For example, while our system correctly scores the phrase "not complex" as being simple, it will incorrectly score the phrase "not *very* complex" as complex because that particular word grouping is currently absent from the dictionary (whereas the word "complex" is present as an indicator of complexity).

Young and Hermann suggest the following flexible rule be applied to solve this problem: "Allow the negation to occur up to 'x' number of words prior to 'complex.' " While this rule solves the problem for the phrase "not very complex," it also adds additional noise in many other cases. This is in part because there are an almost infinite number of possible words that could appear between "not" and "complex," and these words could impact complexity in divergent ways. As Young and Hermann point out, it is simply not feasible to anticipate all of the variations of intervening words. Consider the following examples using just a single intervening word:

(1) It is not only complex; it is emotional.
(2) It is not just complex; it is admirable.
(3) It is not merely complex; it is concerning.

In each case, the flexibility rule would negate the word "complex" due to the preceding word "not." But this would be a misinterpretation of the intent of the statements. The actual intent of these sentences is to use "complex" in a complex fashion. In the example "not *very* complex," the intervening word serves to reinforce the negation; in the example "not *only* complex," the intervening word reinforces the positive assertion of complexity.

Further, we agree with Young and Hermann that a flexibility-based coding scheme becomes progressively less effective as the number of intervening words used by the rule increases. We have used examples above with the smallest number of possible sequence interrupters (one). But if one has a flexible rule that allows for, say, five sequence interrupters between the words "not" and "complex," one increases exponentially the possible meaning-altering words that may occur between the two focal words. The more verbiage that interferes with two complex indicators like "not" and "complex," the less certain we can be about the validity of the algorithm.

It is also worth noting that "not very complex" is actually more complex than "not complex;" but a flexible system treats those two as exactly the same. In actual fact, "not very complex" implies a qualification—that while the thing in question is kind of complex, it is not very complex. A specificity-based system allows us to make more direct distinctions between subtle differences in the likelihood that a particular phrase is probabilistically associated with complexity.

Alternative Word Meanings in Specific and Flexible Systems

We agree with the other symposium members that computer systems are less equipped to handle subtle nuances in language. IC scoring often involves semantic interpretation based on contextual factors. Indeed, as Tetlock et al. (2014) point out, "the structure of words can never be more than a context-dependent proxy" (pp. x). Words can be used as complex in one context and simple in others.

Consider the example of "television debate." This phrase is problematic for our system because the interpretation of "debate" is context dependent. "We are open to debate" would be complex; but "television debate" is merely descriptive and not complex at all. To solve this problem, Young and Hermann propose a flexible rule that evaluates complexity based on a word's part of speech. For "television debates" specifically, the rule would count "debates" as complex, but only if used as a verb. On the surface, this seems like a logical and rather promising approach. And it may in fact be a viable solution, but it seems that it will require more development. Take the following examples:

(1) The two candidates are debating on television.
(2) They debated far too long for my taste.

Neither statement is complex: In both cases, the variation of "debate" is used as a simple descriptor of an event. Yet, according to the flexible rule, "debating" and "debated" are verbs and would thus be counted (inaccurately) as complex.

Thus, while we see the promise of this kind of flexible rule, we think that much development is necessary before it would work effectively to score complexity. We recognize that Young and Hermann were merely giving simplistic examples for illustrative purposes and that such algorithms can become more and more fine-tuned. But our point is that every flexible algorithm, no matter how refined, can potentially have unintended consequences—and the more complicated and flexible, the harder those consequences are to track, measure, and predict. Flexible rules introduce uncertainty.

The Absence of Linguistic Markers in Specific and Flexible Systems

Consider, too, the issue discussed in this symposium of what to do when clear linguistic markers of complexity are absent. The statement "the play was bitter and funny" is clearly complex because bitter and funny are obviously different dimensions applied to the same object; and yet, because it contains no clear linguistic makers of complexity, our system would (incorrectly) score it as simple.

The absence of clear linguistic markers is definitely problematic for our system, and both Tetlock et al. (2014) and Young and Hermann (2014) suggest using sentiment analysis (including algorithms such as *if this sentence contains one positive and one negative word, then it is complex*) to overcome this problem. However, creating lists of "positive" and "negative" dimensions may be in part effective, but it also adds a lot of additional noise.

Consider, for example, the following statement:

Some people said the play was bitter, but actually it is funny.

This statement scores a 1 for IC. But a sentiment analysis using a *one positive word and one negative word* = complex rule would (inaccurately) count it as complex ("bitter" + "funny"). And it seems to us that the possible permutations of the ways positive and negative words can be used in even a single sentence are nearly endless—and that these ways frequently diverge in nonobvious fashion from straightforward sentiment rules. We may be wrong; we are open to the newer developments in this regard (e.g., Wilson, Wiebe, & Hoffmann, 2009) and would love to see them successfully incorporated into an automated system. At present, however, we do not think that it is a certainty that such a system would not add more noise than it gains in precision.[1]

Different Kinds of Complexity

As noted throughout this symposium, complexity is a multifaceted construct. Here we briefly explore some of the issues pertinent to two different distinctions relevant to complexity measurement: (1) integration versus differentiation and (2) elaborative versus dialectical forms of complexity.

Differentiation Versus Integration

In considering different aspects of complexity, we originally focused primarily on the distinction between differentiation and integration. Expanding on this, the additional symposium articles highlighted the usefulness of automated systems—over and above human scoring—in more easily providing tests of the independent and/or overlapping effects of these two aspects of complex thinking. We agree, and indeed we believe part of the added value of our own system is that it provides separate scores for integration and differentiation.

Young and Hermann (2014) also discuss some difficulties with the fact that very few paragraphs actually contain integration. We don't entirely disagree, but we would like to note that the fact IC scoring tends to be skewed towards differentiation does not mean that integration is invalid or not useful. Higher (integrative) scores may be infrequent; but when they occur, they are still meaningful. During an exchange with symposium members, Peter Suedfeld (personal communication) summed up our own view by noting: "I'm always intrigued by the frequent comments that the low proportion of scores above 5 calls for some change to the scoring system. Should we abandon IQ tests and research on IQ because the proportion of scores drops pretty reliably when you get above 140 or so?"

Elaborative Versus Dialectical Forms of Complexity

It is also worth considering more fully—as do both Young and Hermann and Tetlock and colleagues—that there are many other dimensions on which we can discuss complexity besides differentiation/integration. Because these dimensions often have different psychological precursors and consequences (e.g., Conway, Thoemmes, et al., 2008; Conway, Dodds, Towgood, McClure, & Olson, 2011; Tetlock & Tyler, 1996), it seems an important task for moving automated measurements forward to incorporate these additional elements of complexity into computer-based systems.

We ourselves have recently begun automating one of these additional distinctions discussed in prior work and in this symposium—the distinction between dialectical and elaborative forms of complexity. Using our specificity approach, we developed an automated system to parallel human scoring of these constructs, which are scored on the same 1–7 scale as integrative complexity. This

[1] Young and Hermann also discuss, as one of the advantages of a multiple-pass system, the process of *lemmatization*. However, this feature is not unique to multiple-pass systems, and, indeed, our own *Automated IC* system already uses a form of lemmatization.

automated system, like its human-scored counterparts (Conway, Dodds, et al., 2011; Conway, Thoemmes et al., 2008), parses the overall integrative complexity score into its elaborative and dialectical parts.

While the automated system is still in development, early returns have been promising. Using the same trained and untrained data analytic approach as reported for integrative complexity in the main manuscript, we achieved similar correlations for both dialectical complexity (across all nine datasets, average automated-human $r = .40$; for trained data, average $r = .48$; for untrained data, average $r = .35$) and elaborative complexity (across all nine datasets, average $r = .44$; for trained data, average $r = .44$; for untrained data, average $r = .43$). Further, because dialectical and elaborative complexities are sometimes correlated, we performed tests concerning how much these systems discriminated each subconstruct from the other. In short, these tests revealed that, although there was some leakage from dialectical to elaborative and vice versa, the automated system was consistently distinguishing between the two subconstructs (average discriminating $r = .29$; trained = .27; untrained = .29). These data suggest that our preliminary efforts at constructing an automated system for the subconstructs, while not quite as good as the *AutoIC* for integrative complexity, are nonetheless promising.[2]

The Gold Standard

These symposium articles raised some questions concerning what complexity is and the degree to which it can reasonably be measured; even measured by humans. We start with the larger question at the fore of many of these debates: What is the *gold standard* for measuring complexity?

The Irreducible Human Element

All symposium articles have observed, in one way or another, that human judgment cannot directly be removed from the *gold-standard* equation. We agree. Humans will decide what the ultimate standard is; and regardless of the standard that is imposed, humans must serve as the ultimate arbitrators of it, because *any* standard inevitably requires some subjective human judgment to determine if it has been met.

But what *is* the gold standard? In our view, the most reasonable gold standard ought to be correspondence with (1) *expert human scorers* (on as wide a variety of documents as possible) using (2) *prospective* tests. We discuss the importance of each of these in turn.

Expert Human Scorers. Using expert human scorers as an ultimate barometer of validity is directly stated or implied by all the articles in this symposium, so we will not spend time reiterating all the arguments the other articles have made in this regard. Rather, here we discuss some of the issues that were raised in this symposium about how specifically we ought to apply this standard.

In particular, Young and Hermann suggest that the gold standard ought to be a set of documents developed and agreed upon by experts defining the construct. We think this suggestion is a useful starting point for training and testing—but also potentially limiting as an ending point. There is a danger in focusing only on documents or paragraphs on which we can get a set of experts to agree. Consider the analysis Young and Hermann present on 84 paragraphs that their group selected to be clear, representative examples of paragraphs on which IC experts agree. If we

[2] The prototype for auto-scoring elaborative and dialectical complexity is currently available with the *AutoIC* package that can be downloaded for free online—the package scores all three constructs (integrative complexity and the two subconstructs) at the same time.

hand select the easiest paragraphs to score and use that as the gold standard for testing, then it will likely be fairly straightforward to achieve very high correlations for automated systems. But is this a good standard for the effectiveness of the construct? While we think this is an important starting point for measurement, we worry that as an end point—a *gold* standard—it sets the bar too low.

Because (1) most materials are not that easy to score and sometimes even experts do not completely agree on them (we come back to the implications of this problem below), and (2) it is likely that future researchers will not limit the application of an automated construct to only previously selected materials on which experts are likely to agree, we think it is important to not limit our standards only to a set of exemplars that clearly distinguish high- and low-complexity statements. So while easy-to-score exemplars are useful for defining the construct—and we agree with Young and Hermann that more could be done for IC in this regard—we think using those documents ought not to be the only thing involved in testing an automated system.

Conceptually, we think the ultimate standard for testing any automated system ought to be its correspondence on *all* possible paragraphs—across a wide variety of materials—with the collective *average view* the world's leading experts on complexity. Because such experts cannot possibly code all possible paragraphs in even a typical project—much less the entire world's statements—we view correspondence with scorers *trained by experts* as a go-between proxy of truly expert scoring. This method allows for the development of automated systems on a wider array of existing materials that includes both easy-to-score and more difficult-to-score materials. So while the typical undergraduate coders in our lab are not experts in the sense that Peter Suedfeld and Phil Tetlock are experts, we view the correspondence of an automated system with these typical human scorers to be a vital and appropriate cog in the advancement of IC research.

Prospective Versus Ad Hoc Tests

Correspondence with human scorers is thus an irreducible part of any formula for designing a gold standard. However, we also think another point (only briefly mentioned in our original article and alluded to in Young and Hermann's article) is worth elaborating on here. In particular, in our target article, we performed *ad hoc* tests on data we had "trained" our system on, as well as *prospective* tests on data we had not "trained" on. We emphasized in our original paper that the prospective, "untrained" tests were more important as markers of the validity of our system. Here, we would like to elaborate on why this is important and what it means for the question of proper tests for automated systems.

To illustrate the importance of the prospective/ad hoc distinction, we performed some additional analyses. We focused these analyses on the dataset that our *AutoIC* system had the most difficult time scoring—the Nixon/Kennedy debates (for the *AutoIC*, $r = .18$). Using our specificity approach, we created an *ad hoc* system designed to accurately score complexity only in those debates. We were successful: The correlation between our *ad hoc AutoIC* system and human-scored IC for the Nixon-Kennedy debates was $r = .87$. However, when we took this successful ad hoc system to the other eight datasets, the results were less than overwhelming: The average correlation in those additional datasets was $r = .26$. This simple test illustrates that, while it is fairly easy to get high correlations on any one dataset (creating this new *ad hoc* system took us only one afternoon of work), the real test of any system is its ability to predict human IC on multiple different kinds of datasets in a prospective fashion. Our *AutoIC* was designed from a deductive approach with this specific goal in mind. This means that, while its correlations are not as high as they could be had we only attempted to gain high correlations on our existing data—and granting that there will always be variability in correspondence across datasets—it increases the odds that the *AutoIC* will continue to achieve moderate correlations in the future.

Beyond Mere Correspondence: Other Possible Markers of Complexity

In defining a gold standard and testing our own system, we have focused our efforts almost entirely on correspondence with human scorers, and (as we have argued here) we believe this is ultimately the most effective validity marker. However, as our colleagues have noted correctly, this is still only one of many ways we can consider *success* in measuring this difficult construct. A different approach, raised by all the articles in the symposium, would be to show the predictive ability of the measurement in actual research studies.

Clearly this approach is valid; no system will ever be scientifically useful unless it predicts (or is predicted by) something. It is worth noting, however—as both Tetlock et al. (2014) and Young and Hermann (2014) alluded to—that not all predictive validity tests of this sort are equal in their usefulness for determining if a measurement is truly about complexity. To illustrate, we think it is useful to divide these sorts of tests into three separate categories.

First-order effects: Tautological tests. One category involves tests that any system—human or computer—should reasonably pass if it were truly measuring complexity. For example, our lab plans to run some studies where we tell participants what complexity is and ask them to write essays using either really complex or really simple language. Assuming participants read and encoded the directions correctly, one would assume that a failure to find differences between the simple and complex groups in this study would suggest the measurement was not picking up very well on real-time complexity.

Second-order effects: Tests implied by the measurement. There also exists a category of tests that, while not quite tautological in nature, would still be expected (on average) to exhibit effects on a complexity measurement. For example, in our own tests, we expected that more descriptive documents ought to be lower in complexity than more apologetic/argumentative kinds of documents (they were). This does not necessarily follow directly from the definition of complexity in the same way as first-order effects, but it would still be surprising not to find it because it seems intuitively clear that descriptive statements should on average be less complex than fully argued statements. Similarly, one might argue that a failure to find an increase in complexity under conditions implied by the *Value Pluralism Model* (e.g., Tetlock, 1986) would constitute a failure to find something directly implicated by the measurement of complexity.

Third-order effects: Tests expected by a specific theory. A third category of tests involve not effects that are implied by the measurement of complexity itself but rather effects that some specific theory of psychology would predict on other grounds. Such tests are useful in so much as they can show the meaningfulness of the measurement to capture a valid finding, but they are also less useful as direct barometers of the measurement of complexity. Take, for example, the hypothesis that international crises involving drops in complexity on both sides are especially likely to lead to war (e.g., Suedfeld, Tetlock, & Ramirez, 1977). This is based in a specific theory of the psychology of violence. A failure to find an effect could mean that the measurement is invalid, but it could also mean that the theory is wrong. Furthermore, a success at finding an effect could mean that the measurement is valid, or it could mean that it measures something besides complexity, and it is that other variable that is associated with war/peace outcomes.

This latter problem is particularly in evidence for computer scoring, where it is highly possible that word dictionaries overlap with multiple categories. For example, it is likely that integration words (“put together”; “integrate”) are at least mildly positively correlated with peaceful words.

It is also worth noting that, as Tetlock et al. (2014) describe in detail, often competing theories or data exist—indeed, some theories might predict that complexity would go up prior to violence due to increased processing demands (e.g., Hermann & Sakiev, 2011)—and this complicates things further. Thus, while these third-order findings are useful in some sense for showing a measurement

is capturing something beyond mere chance, they should not be treated as perfect barometers of the exact nature of the measurement itself.

A note on replicating effects found for human scoring. Finally, we might consider replicating effects found on human scoring as a barometer for the goodness of a computer measurement. Young and Hermann, for example, note the overlap with human and computer scoring concerning the U.S. Presidents' campaign rhetoric (Hermann, Sakiev, & Smith, 2010; Thoemmes & Conway, 2007). We think this strategy very useful, but we also caution that the same set of difficulties laid out for third-order effects above also applies here: Because a lack of correspondence between types of effects could occur for many possible reasons, one should be cautious of overinterpreting it.

The (Sometimes) Difficulty in Human Scoring of Complexity and What It Means

Complexity is clearly a difficult construct to score. Most linguistic constructs are scored based on their semantic meaning alone, so that it is often easier to construct a set of automated semantic words directly related to it (e.g., power). Complexity is not like that. It is—as our colleagues pointed out—very "fuzzy." Although, in the broader panorama of psychology, such fuzziness is not unique to complexity research (see, e.g., Pelham & Blanton, 2012; Tweed, Conway, & Ryder, 1999)—it still raises legitimate questions about what it means for automating the construct.

Correlations Among Human Scorers Versus the Computer

For example, it is possible that one of the reasons it is difficult to get a computer to agree with a human is that it is sometimes difficult to get humans to agree with each other. This implies a testable hypothesis: Namely, that on materials for which it is more difficult to get humans to agree, computers ought to have a harder time agreeing with humans as well.

We used the data from the focal paper to provide a set of crude tests of this hypothesis. First, we looked at the average correlation between scorers across all our data that had more than one scorer (this excludes the practice materials and coding test, which were scored primarily by one expert). We then evaluated the degree that higher correlations among humans predicted higher computer-human correspondence. There was a modest tendency for higher levels of human agreement to predict better human-computer correspondence ($r[7] = .22$). However, this figure is somewhat misleading as an indicator of the difficulty of achieving human agreement because for two of the seven datasets in question, our lab recoded the set due to lower initial levels of agreement. To account for this, we returned to those datasets and computed the original level of agreement. Using original levels of agreement as a marker of the difficulty of achieving human agreement, there was a much larger tendency for higher levels of human agreement to predict better human-computer correspondence ($r[7] = .75$). Finally, we created a dummy-coded variable indicating whether or not we had to rescore data (=0) or not (=1). This variable was a strong predictor of human-computer correspondence, with a much larger tendency for easy-to-score materials to predict better human-computer correspondence ($r[7] = .87$). This reflects the fact that the two datasets that were most challenging for our lab in terms of achieving human agreement—the Nixon/Kennedy set and the Bush/Kerry set—were also the sets that showed the lowest correspondence between human scoring and our computer scoring.

In short, this set of analyses provides empirical support for what our colleagues have argued—that part of the problem in computer scoring is with the difficulty of human scoring. As humans have less difficulty scoring materials themselves, computers are better at corresponding with them.

Lack of Absolute Agreement and Its Implications

This sometimes difficulty in coding complexity raises several additional questions concerning the viability of the construct and its translation to automated systems. For example, does the fact that

it is difficult to gain absolute levels of agreement undermine the value of the construct? Young and Hermann and Tetlock et al. both note that one can get a high correlation among human scorers when absolute levels of agreement are low. This is clearly true.

Does lack of agreement imply scientific invalidity? And yet it does not follow from this fact that the correlation is not a relevant marker of scientific validity. Indeed, it is worth noting that often absolute *dis*agreement is over a 1-point difference in a paragraph (as is the case in the example used by our colleagues). The basic question over absolute agreement versus correlation can be stated as follows: Are those small disagreements more meaningful than the fact that both coders assigned scores in the same general direction? Consider two paragraphs from our colleagues' hypothetical set: Paragraph 1 was scored a 1 by the two coders. Paragraph 18 was scored a 6 and a 7, respectively, by the same two coders. What is more important—the fact that absolute agreement for those two paragraphs is only at 50%, or the fact that both coders agree that the first paragraph is really low in complexity and the second is really high? A correlation accurately captures the latter—the general agreement of scorers of the general complexity of paragraphs. Given that the task is one where no actual, true marker of complexity exists, we find that correlations are a very useful measure.

Distinguishing the conceptual definition from its application to measurement. Tetlock et al. (2014) make a useful distinction that is worth revisiting briefly here. Namely, they note that there is a difference between the question "do experts in this field agree generally what this construct is?" and "do experts in this field agree on specific instances of measurement?" They suggest that to progress more fully, we need to move forward collectively on both questions.

We agree; but, while we think there is some disagreement about both types of questions, the problem is much more pronounced for specific measurements than about the definition of the construct. A larger glance at the history of complexity research shows that, from a definitional standpoint, almost all forms of complexity contain the concept of *multidimensional differentiation*—that is true for self-report complexity-relevant measurements of all kinds (e.g., Bieri et al., 1966; Fletcher, Danilovics, Fernandez, Peterson, & Reeder, 1986; Neuberg & Newsom, 1993), for self-complexity measurements (Linville, 1982; Scott, 1969), and for all of the measurements for scoring transcribed materials discussed in this symposium. It is also true for different *facets* of differentiation, such as elaborative and dialectical complexity. Although all these forms/facets of measurement are very different (see, e.g., Vannoy, 1965), in each case, something with multiple dimensions is considered more complex than something with only one dimension.

That is not to say that the term "complexity" is always used to mean the same thing: We ourselves have noted in the past that it is not (Conway et al., 2008), and we agree that an important task is to continue to define the construct (or at least be clear in our research about which definition we are using). And we acknowledge that our own preferred conceptualization—which includes integration—is common across some, but not all, conceptualizations. Nevertheless, it is worth noting that much of the complexity-relevant fuzziness (and resulting measurement difficulty) we are discussing in this symposium exists more in the application of the construct to difficult cases than it does in how we view the construct writ large. Coders may disagree over whether or not, in a given paragraph, a politician meant economic and foreign policy consequences to be truly differentiated; but coders do not disagree over whether or not the statement "this problem has only one possible solution" and the statement "this problem has multiple possible solutions along the completely different dimensions of reason (on the one hand) and emotion (on the other)" differ in complexity. That is because expert coders generally agree on what the construct is and can thus generally agree on its application in its most clear instantiations. While, as argued above, we think it important to delve into the difficult cases in order to improve measurement, it is also important to keep in mind that the construct is not nearly as *fuzzy* in its generally accepted definitional meaning as it is in its application.

Evidence That Complexity Measurements May Be More Coherent Than They Seem

Although we acknowledge, along with our colleagues, that there is a sense that the reported correspondence between computers and humans on complexity probably overestimates the actual degree of correspondence, we also think there are a couple of ways that the reported correlations between various complexity measurements might actually underestimate the correspondence as well. Below, we present two sets of additional data to illustrate.

Increasing power: The unit of analysis. For example, consider that in our target article we likely presented the most conservative unit of analysis. Combining smaller units (in our case, paragraphs) into larger categories (documents) often increases the acuity and power of measures because the loss of numerical power is more than compensated for by the reliability gained from combining multiple instantiations of like categories (see, e.g., Conway, Dodds, et al., 2011; Lashley & Kenny, 1998).

To evaluate this possibility in our own data, for all of our datasets which allowed it, we created person-level (presidential elections) or book-level (ancient Christian writings) summary scores for each document (for both AutoIC and human-scored IC). Further, as reported in our target article, we performed similar correlations at the level of the summary dataset. As we used increasingly larger units of analysis, human-computer correspondence increased ($r = .46$ for paragraph level; $r = .55$ for person/book level; $r = .82$ for document level). Thus, although we chose to use the more conservative tests in our target article, it is possible that the gap between human and computer correspondence is not as large as it seems.

Correspondence among different computer systems. It is also worth considering the measurement of complexity beyond integrative complexity (which is only one of many different possible measurements of complexity). Historically, human-scored measurements of complexity are not highly correlated with each other (e.g., Vannoy, 1965; Suedfeld, Frisch, Hermann, & Mandel, under review), although they sometimes predict similar things (e.g., Vannoy, 1965). There are many potential reasons for this, including that they may be measuring complexity on different domains, of different types, or with a somewhat different conceptual focus (see, e.g., Conway, Thoemmes, et al., 2008; Tetlock & Tyler, 1996). Here we would like to suggest a different reason: Namely, that part of the lack of correlation between two conceptually related complexity measurements may occur because they were designed to be used differently and at a different methodological level. We illustrate by presenting some additional data on the automated Conceptual Complexity measurement that represents a collaborative effort between our lab and Hermann's group of researchers.

One of the issues with comparing Conceptual Complexity and Integrative Complexity is that the two measurements were designed around different units of analysis—Conceptual Complexity was designed to be assessed using longer, topic-focused units, while the focus of Integrative Complexity is shorter paragraphs of a sentence or two. To obtain a three-way comparison between Automated Conceptual Complexity (CC), Automated Integrative Complexity (AutoIC), and Human-Scored Integrative Complexity (HumanIC), we (in conjunction with Hermann's group) selected a dataset that contained fairly long paragraphs—the religious writings from the original paper. Hermann's group scored all the paragraphs from the religious writings for CC. Some of those paragraphs ended up giving inadequate data for CC because they were too short. As such, analyses were performed both with and without these data.

Consistent with the idea that part of the lack of overlap between the constructs involves different optimal units of measurement, the correlation between CC and AutoIC was greater when these shorter paragraphs were excluded ($r[124] = .21$) than when they were included ($r[168] = .13$). The same pattern held when comparing CC and HumanIC (excluding short paragraphs $r[124] = .24$; including short paragraphs $r[168] = .19$).

Further, Hermann's group also scored CC for three presidential campaigns (Obama/McCain, Nixon/Kennedy, and Bush/Kerry) from the main paper that used the IC units of analysis. Since these units were generally too short to score for CC, the paragraphs were combined by assigning pairs of

identical Automated IC scores and scoring CC for the combined paragraphs. Using this method, correlations between CC and AutoIC showed a similar relationship as the religion data for longer (and more appropriate for CC scoring) paragraphs, (r[49] = .23).

Overall, these correlations between CC and both HumanIC and AutoIC are fairly similar to prior reported correlations between CC and HumanIC (Suedfeld, Frisch, et al., under review). But the fact that the correlations are higher on longer materials suggests that at least part of the reason the two systems do not correspond in practice is because they were designed around different units of analysis.

Concluding Thoughts: Mapping a Way Forward

In spite of the difficulties of automating complexity, computers clearly do offer some advantages over humans—and all symposium members appear to agree on this fact. Because they can score substantially more data, to fully take advantage of the exponentially expanding world of linguistic data, automated systems are incredibly important. Human-scored systems were, in a sense, developed in a different era where the amount of available data to score was not large, and therefore slower methods of scoring could still score quite a bit of the relevant data. Today that is not the case. Therefore, we think it is important, moving forward, to develop automated systems and to do them as best we can.

So where does all this leave us? Which system should we use, or should we use a combination of systems? We think the answer to these questions depends in part on whether we are talking about measuring complexity with a large panoramic lens or the more specific question of accurately approximating human scoring of integrative complexity in particular. We close with a brief discussion of each.

The Big Picture: No Need to Choose Between Systems

Of the available automated systems relevant to complexity, which system do we recommend using? In the bigger picture, we agree with our colleagues that there is no necessary reason to *choose* between systems. It is possible that such systems can continue to coexist side-by-side and, as such, offer useful independent insights into the nature of complex rhetoric. There is indeed value in having such independent measurements: To the degree that they measure somewhat different aspects of complexity, they can help us better understand the overarching relation of complexity to human psychology by the principle of *triangulation* (see, e.g., Conway, Houck, & Gornick, in press; Kitayama, Conway, Pietromonaco, Park, & Plaut, 2010). Removing such variability in favor of a single approach or system might in fact stagnate science by unnecessarily restricting the possible dimensions or facets of a clearly multifaceted construct.

The Accurate Measurement of Integrative Complexity Is an Empirical Question

Of course, we are naturally biased towards integrative complexity approaches for the reasons specified in our target article; and it is still worth considering the more specific goal of measuring the integrative complexity construct itself as accurately as possible. Given that goal, to what degree do we see the LTA/flexibility approach as a viable alternative to either be used instead of, or in combination with, our AutoIC system? And we would like to emphasize that our answer to this question ought to be *based on empirical evidence*. This symposium has laid out some pretty specific empirical criteria—criteria that, for the most part, symposium members agree on—for judging the validity of automated scoring in general and integrative complexity in particular. We are certainly open to the usefulness of LTA/flexibility approaches to coding complexity and see that they have many potential advantages. However, to us this is primarily an empirical question. And, at this point, there is not much empirical evidence from flexibility/LTA approaches with respect to correspondence

with integrative complexity. As Young and Hermann note, the *CIC* system was left in a fairly undeveloped state and, as a result, does not have much empirical evidence in terms of prospective human tests with human scorers.

However, with that caveat, it is worth considering that, since the flexibility and specificity approaches have potentially offsetting strengths and weaknesses, some sort of *combination* strategy might ultimately prove more effective than either approach alone. If that were the case, then it suggests some kind of "super-system"—such as that mentioned by Young and Hermann—might be developed using the best parts of both systems. We agree and would be excited to work on such a system, provided we can empirically justify its validity. One possibility moving forward would involve testing multiple systems on prospective data, both as independent and additive predictors, with the goal of seeing if a combined system predicts more variance in human scoring than each independent system. If it does, then that would serve as an impetus for perhaps creating a larger system for measuring integrative complexity out of the existing ones.

REFERENCES

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

Conway, L. G., III, Dodds, D., Towgood, K. H., 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, Houck, S. C., & Gornick, L. J. (in press). Regional differences in individualism and why they matter. In J. Rentfrow (Ed.), *Psychological Geography*. Washington, DC: American Psychological Association.

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

Fletcher, G. J. O., Danilovics, P., Fernandez, G., Peterson, D., & Reeder, G. D. (1986). Attributional complexity: An individual differences measure. *Journal of Personality and Social Psychology, 51*, 875–884.

Hermann, M., & Sakiev, A. (2011). Leadership, terrorism, and the use of violence. *Dynamics of Asymmetric Conflict, 4*, 126–134.

Hermann, M. G., Sakiev, A., & Smith, M. (2010) Governance in context: Understanding the ingredients of political leadership. Paper presented at the annual meeting of the American Political Science Association, Washington, DC.

Kitayama, S., Conway, L. G., III, Pietromonaco, P. R., Park, H., & Plaut, V. C. (2010). Ethos of independence across regions in the United States: The production-adoption model of cultural change. *American Psychologist, 65*, 559–574.

Lashley, B. R., & Kenny, D. A. (1998). Power estimation in social relations analyses. *Psychological Methods, 3*, 328–338.

Linville, P. W. (1982). The complexity-extremity effect and age-based stereotyping. *Journal of Personality and Social Psychology, 42*, 193–211.

Neuberg, S. N., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simple structure. *Journal of Personality and Social Psychology, 65*, 113–131.

Pelham, B. W., & Blanton, H. (2012). *Conducting research in psychology: Measuring the weight of smoke* (4th ed.). Belmont, CA: Wadsworth/Thompson Learning.

Scott, W. A. (1969). Structure of natural cognitions. *Journal of Personality and Social Psychology, 12*, 261–278.

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., & Tetlock, P. E. (2014). Integrative complexity at forty: Steps toward resolving the scoring dilemma. *Political Psychology, 35*, 597–601.

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

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

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

Tetlock, P., Emlen Metz, S., Scott, S., & Suedfeld, P. (2014). Integrative complexity coding raises integratively complex issues. *Political Psychology, 35*, 625–634.

Thoemmes, F., & Conway, L. G., III (2007). Integrative complexity of 41 US presidents. *Political Psychology, 24*, 781–801.

Tweed, R., Conway, L. G., III, & Ryder, A. G. (1999). The target is straw or the arrow is crooked. *American Psychologist, 54*, 837–838.

Vannoy, J. S. (1965). Generality of cognitive complexity-simplicity as a personality construct. *Journal of Personality and Social Psychology, 2*, 385–396.

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

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