Appendix to "Complexity, Professionalism, and Text Borrowing in State Legislatures"

August 11, 2020

Contents

1 Descriptive Statistics and Alternate Model Specifications

Table A1: Descriptive Statistics

Variable	N	Mean	Std. Dev.	Min	Max
Similarity Score	530	61.87	22.84	0	99.22
Alignment Score	530	45.72	23.11	0	100
Issue Complexity	10,254	98.81	4.28	93.2	108.3
Staff Expenditures	10,174	6.47	6.81	0.48	55.23
Salary	10,174	53.89	46.21	0	254.9
Session Length	9,327	145.34	86.88	36	549.5
Term Limits	10,246	0.16	0.37	0	1
Per Capita Income	10,207	3.63	0.64	1.92	5.95
Ideological Distance	10,254	0.24	0.17	0	0.83
Government Ideology	10,208	0.51	0.24	0	0.95
Border	10,254	0.14	0.24	0	1
Word Count	10,254	6.74	1.09	4.92	8.15
Time	10,254	8.83	7.17	0	28
Order	527	14.17	12.75	0	49

Figure A1: Policy Rankings by Average Complexity

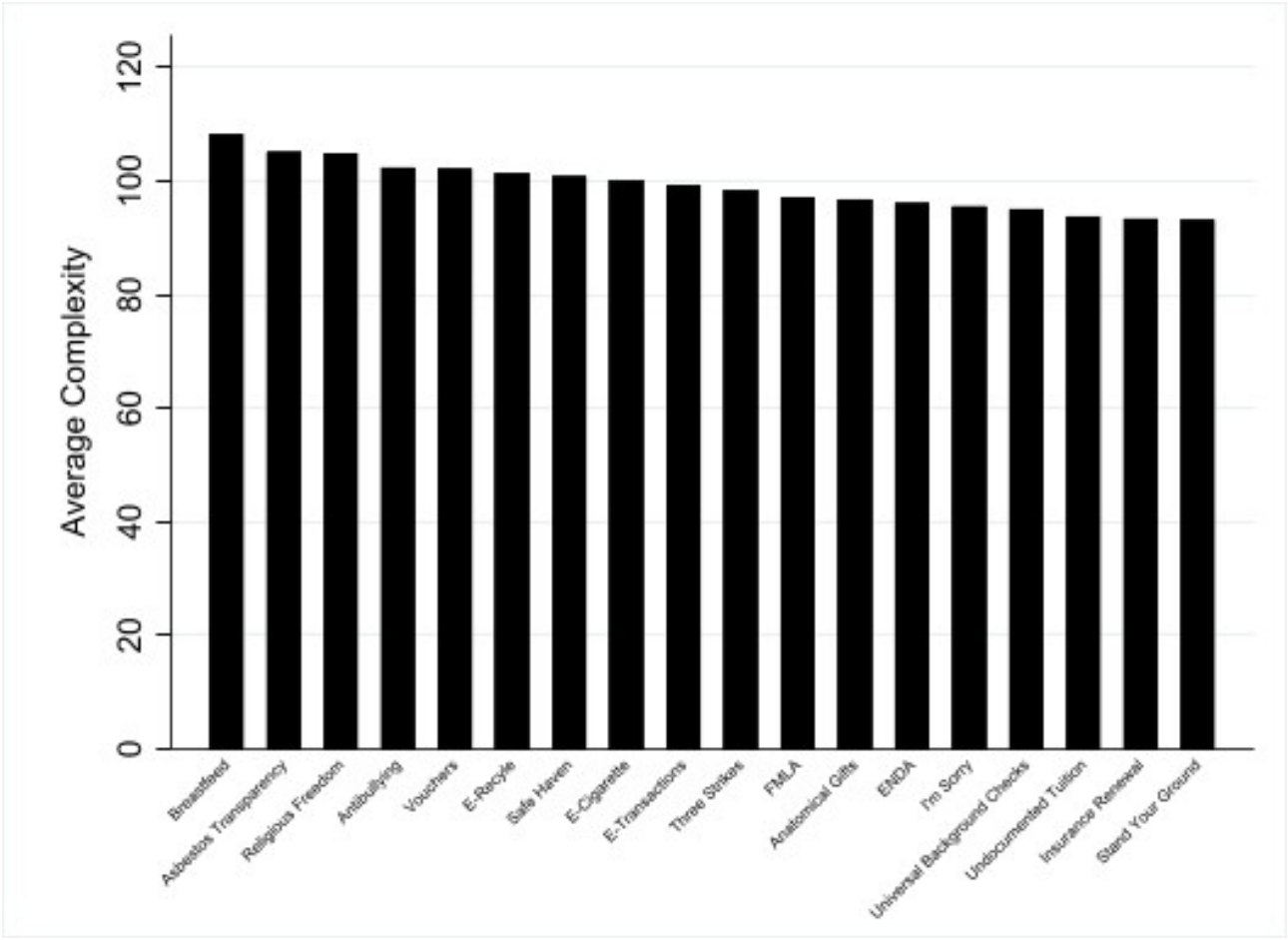


Table A2: Selection Model of Policy Language Diffusion Using Logit Transformed Second Stage DV

	(1)	(2)
	Stage 2: DV = Transformed Sim Score	
Staff Expenditures	-0.63	-20.42*
	(0.58)	(7.80)
Salary	-0.04	-0.04
	(0.15)	(0.15)
Session Length	-0.11	-0.10
	(0.06)	(0.06)
Term Limits	18.74*	17.71*
	(7.75)	(7.69)
Model Legislation	37.30*	37.84*
	(6.73)	(6.76)
Complexity	-4.81*	-6.11*
	(0.81)	(1.02)
Time	0.63	0.73
	(0.78)	(0.79)
Order	2.48*	2.49*
	(0.50)	(0.50)
Word Count (log)	41.52*	41.78*
	(4.38)	(4.37)
Expend X Complex		0.20*
		(0.08)
Constant	270.45*	396.87*
	(78.58)	(100.50)
	Stage 1: DV = Adopt	
Ideological Distance	-1.19*	-1.20*
	(0.20)	(0.20)
Government Ideology	-0.55*	-0.55*
	(0.18)	(0.18)
Model Legislation	0.11*	0.11*
	(0.04)	(0.04)
Border	1.60*	1.60*
	(0.11)	(0.11)
Per Capita Income	0.25*	0.25*
	(0.05)	(0.05)
Time	-0.05*	-0.05*
	(0.01)	(0.01)
Time^2	0.00*	0.00*
	(0.00)	(0.00)
Constant	-2.20*	-2.20*
	(0.23)	(0.23)
ρ	-0.17	-0.17
Inverse Mills Ratio	-15.33	-15.21
	(11.31)	(11.42)
N	10169	10169
BIC	7748.19	7755.66

Note: $p \leq .05$. Robust clustered standard errors are in parentheses. Significance tests are two-tailed.

Table A3: Selection Model of Policy Language Diffusion Using Smith-Waterman Alignment Scores

	(1)	(2)
	Stage 2: DV = Alignment Scores	
Staff Expenditures	-0.25*	-3.38*
	(0.09)	(0.84)
Salary	0.02	0.02
	(0.03)	(0.03)
Session Length	-0.04*	-0.04*
	(0.01)	(0.01)
Term Limits	2.60	2.42
	(1.62)	(1.62)
Model Legislation	7.65*	7.63*
	(1.55)	(1.57)
Complexity	-0.52*	-0.74*
	(0.15)	(0.18)
Time	0.35*	0.36*
	(0.17)	(0.17)
Order	0.41*	0.41*
	(0.10)	(0.10)
Word Count (log)	9.02*	9.02*
	(0.82)	(0.81)
Expend X Complex		0.03*
		(0.01)
Constant	40.01*	61.74*
	(15.55)	(17.24)
	Stage 1: DV = Adopt	
Ideological Distance	-1.11*	-1.11*
	(0.19)	(0.19)
Governemnt Ideology	-0.49*	-0.49*
	(0.17)	(0.17)
Model Legislation	0.07	0.07
	(0.04)	(0.04)
Border	1.56*	1.56*
	(0.11)	(0.11)
Per Capita Income	0.27*	0.27*
	(0.05)	(0.05)
Time	-0.07*	-0.07*
	(0.01)	(0.01)
Time^2	0.00*	0.00*
	(0.00)	(0.00)
Constant	-2.19*	-2.19*
	(0.22)	(0.22)
ρ	-0.24	-0.24
Inverse Mills Ratio	-4.61*	-4.55*
	(1.79)	(1.79)
N	10169	10169
BIC	7748.19	7755.66

Note: $p \leq .05$. Robust clustered standard errors are in parentheses. Significance tests are two-tailed.

2 Validating the Similarity Measure

In this section, we provide results of validation checks for our measure of text borrowing, the cosine similarity score. We begin by illustrating the similarity score with substantive examples from our data set. The example shows three versions of an I'm Sorry Law–Colorado (2003), Iowa (2006), and Nebraska (2007). Nebraska's law is highly similar to Colorado's, with a similarity score of 91.9. A similarity score this high puts Nebraska's bill in the 90th percentile of similarity. The figure shows the copied text highlighted in gray and key policy differences in bold. Nebraska's law is nearly word-for-word the same as Colorado's law, with one key policy difference: Colorado includes admissions of fault as inadmissible in medical malpractice suits, but Nebraska does not adding the sentence "A statement of fault...shall be admissible" onto the block of otherwise copied text. Iowa's law is not very similar to Colorado's, with a similarity score of 33.64, which is about the 10th percentile in similarity. Although the law mostly does the same thing but using different language, Iowa also adds protections for the plaintiff and their relatives. These sorts of examples are common in the dataset; the cosine similarity method is picking up on how much text is actually borrowed and the similarity score drops not just when synonyms are used but when key policy provisions are added or subtracted.

Figure A2: Similarity Example

Colorado's 2003 I'm Sorry Law:
In any civil action brought by an alleged victim of an unanticipated outcome of medical care, or in any arbitration proceeding related to such civil action, any and all statements, affirmations, gestures, or conduct expressing apology, **fault**, sympathy, commiseration, condolence, compassion, or a general sense of benevolence which are made by a health care provider or an employee of a health care provider to the alleged victim, a relative of the alleged victim, or a representative of the alleged victim and which relate to the discomfort, pain, suffering, injury, or death of the alleged victim as the result of the unanticipated outcome of medical care shall be inadmissible as evidence of an admission of liability or as evidence of an admission against interest.

Nebraska's 2007 I'm Sorry Law (Sim Score w/ CO = 91.9):
In any civil action brought by an alleged victim of an unanticipated outcome of medical care, or in any arbitration proceeding related to such civil action, any and all statements, affirmations, gestures, or conduct expressing apology, sympathy, commiseration, condolence, compassion, or a general sense of benevolence which are made by a health care provider or an employee of a health care provider to the alleged victim, a relative of the alleged victim, or a representative of the alleged victim and which relate to the discomfort, pain, suffering, injury, or death of the alleged victim as a result of the unanticipated outcome of medical care shall be inadmissible as evidence of an admission of liability or as evidence of an admission against interest. **A statement of fault which is otherwise admissible and is part of or in addition to any such communication shall be admissible.**

Iowa's 2006 I'm Sorry Law (Sim Score w/ CO = 33.64):
In any civil action for professional negligence, personal injury, or wrongful death or in any arbitration proceeding for professional negligence, personal injury, or wrongful death against a person in a profession regulated by one of the boards listed in section 272C.1 … based upon the alleged negligence in the practice of that profession or occupation, that portion of a statement, affirmation, gesture, or conduct expressing sorrow, sympathy, commiseration, condolence, compassion, or a general sense of benevolence that was made by the person to the plaintiff, relative of the plaintiff, or decision maker for the plaintiff that relates to the discomfort, pain, suffering, injury, or death of the plaintiff as a result of an alleged breach of the applicable standard of care is inadmissible as evidence. **Any response by the plaintiff, relative of the plaintiff, or decision maker for the plaintiff to such statement, affirmation, gesture, or conduct is similarly inadmissible as evidence.**

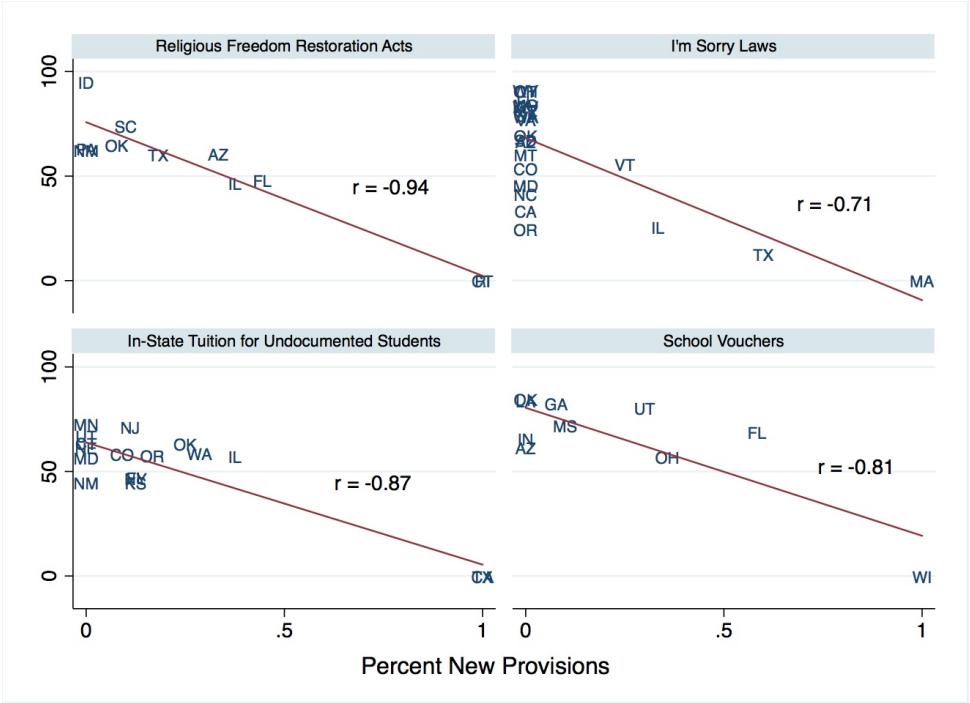
2.1 New Provisions

Ideally, the cosine similarity measure, as a proxy for reinvention, would do a good job of detecting the addition of new policy provisions as the policy diffuses. Classic studies of reinvention note that later adopters build upon the policy by adding provisions to broaden the policy or increase its scope (Clark 1985; Glick & Hays 1991; Hays 1996; Mooney & Lee 1995). The addition of new provisions may increase or decrease its comprehensiveness (Hays 1996), permissiveness (Mooney & Lee 1995), or stringency (Carley & Miller 2012) depending on the specific policy being studied, but across all policies the change in the scope of the policy as it diffuses is dependent on states adding more provisions to the policy.

Our cosine measure, if it measures reinvention and distinguishes reinvention from simple copying, should be strongly negatively related to the addition of more policy provisions. As states adopt new provisions, and craft language to implement those provisions, they should be less similar to previous adopters. To check this, we selected four of our policies and coded for the number of provisions present across bills. We identified 19 unique provisions on across Religious Freedom Restoration Acts, 11 across I'm Sorry Laws, 23 across bills that allow undocumented students to receive in-state tuition, and 34 across school voucher programs. A list of these provisions and which state has each is provided in an spreadsheet provided in the online Supplemental Material for this paper. For each state adoption, we calculated the number of new provisions added over the total number of provisions in the bill as a measure of how much reinvention was occurring relative to how many provisions are actually in the bill.

We plot the percent new provisions on the x-axis and the associated similarity score on the y-axis for each policy area in Figure A3. There is a strong negative relationship between the addition of new provisions and similarity scores, as expected. States that added new provisions were on average less similar to previous adopters than states that did not add many new provisions. The strongest of these is on Religious Freedom Restoration Acts; the weakest case is I'm Sorry Laws where the association is still strongly negative (r = -0.71) but there is substantial noise among adopters that added no new provisions. Some of these states were highly similar in their language to previous adopters and added no new provisions, while others

Figure A3: Correlation Between Addition of New Provisions and Cosine Similarity Scores



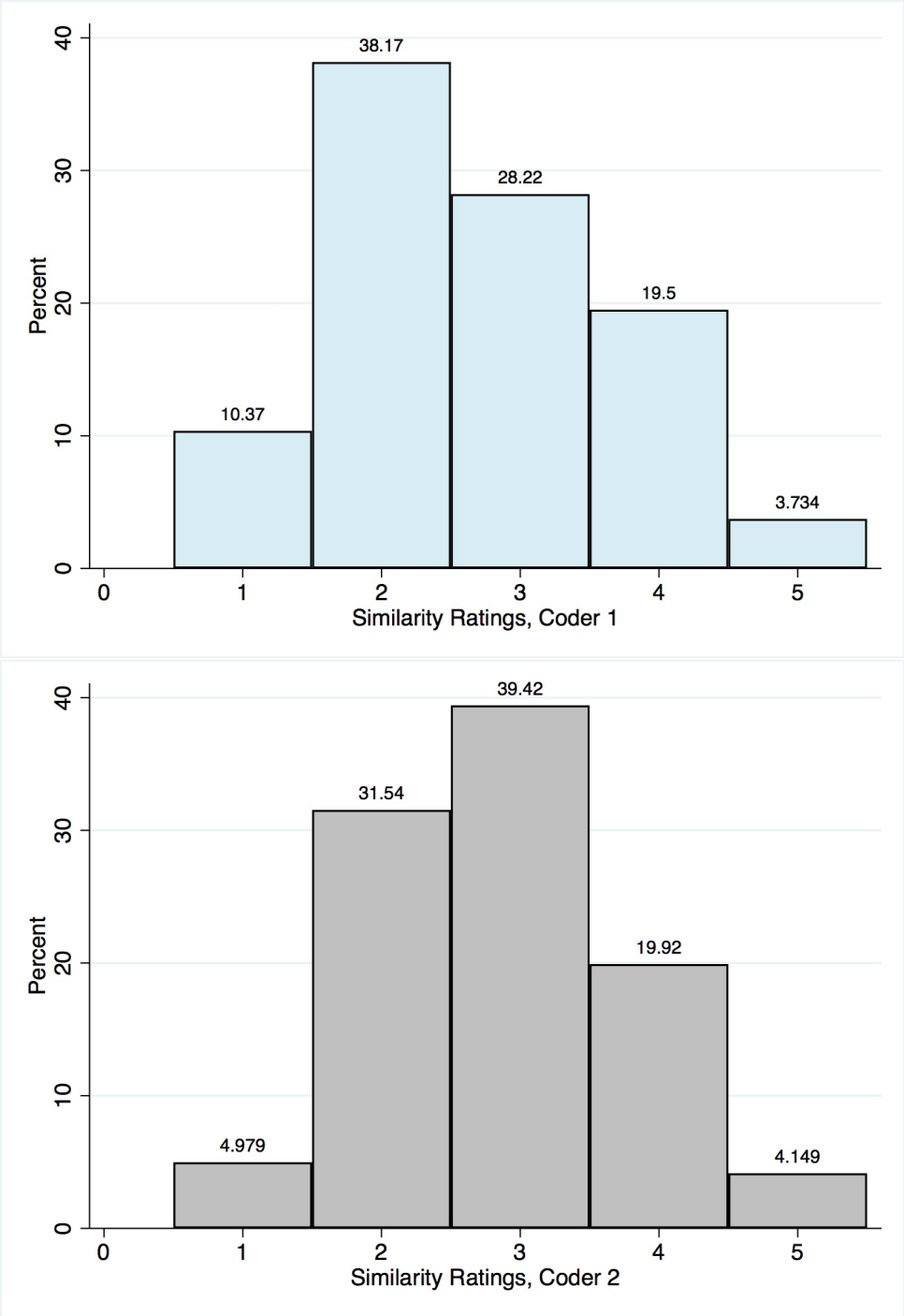
used more unique language but still added no new provisions. It is important to note that I'm Sorry Laws have substantially fewer provisions than the other laws, indicating that states may have been content with adopting the few provisions associated with the policy and tinkering with language as desired. The cosine measure checks out against the theoretical definition of reinvention.

2.2 Human Coding

We also attempted to verify the cosine measure against human coders. Specifically, we asked human raters to code how similar two bills in the same policy area were to one another. We took a random sample of 250 bill pairs and asked student research assistants to rate each pair for similarity. We asked "To what extent do these bills use the same words and phrases." Raters were given five options: 1) A nearly exact word-for-word and phrase-for-phrase match, 2) many words and phrases in common, 3) some words and phrases in common, 4) a few words but no phrases in common, 5) no words or phrases in common. Two raters coded each bill pair, and we ended up with 241 useable responses in which both raters coded the bill pair without any errors. The frequency of response in each category for each rater is depicted in the histograms

in Figure A4.

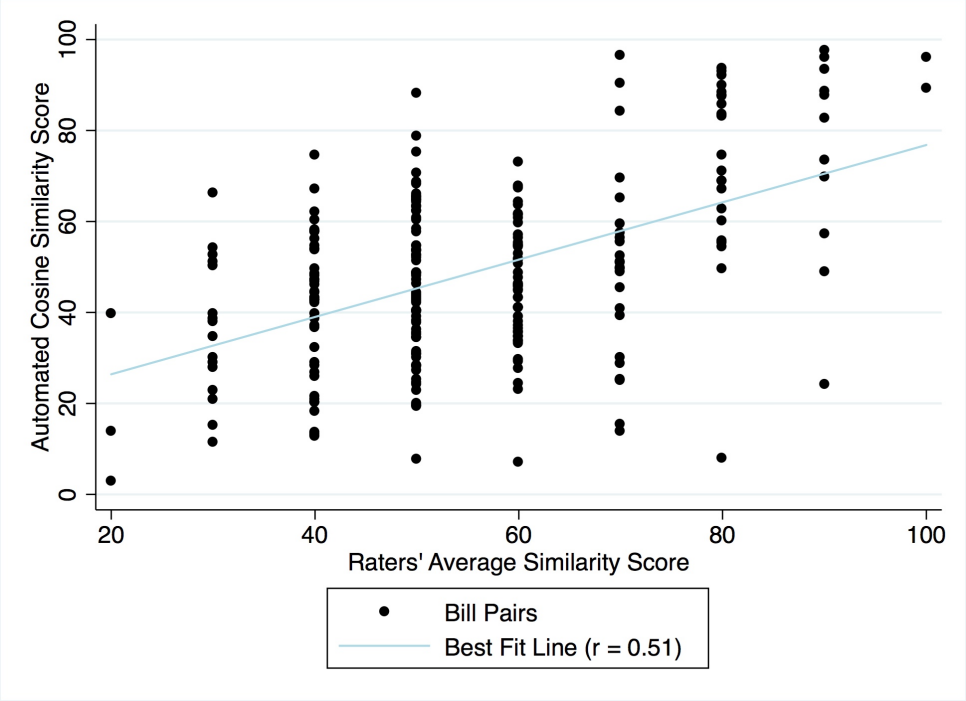
Figure A4: Human Rating of Bill Pair Similarity



Although the raters coded similar percentages of pairs into each of the five categories, there was mixed agreement on which pairs belonged in which category. The raters placed the same pair in the same category 39% of the time. That level of agreement is not simply due to random chance ($p = 0.03$), but it is only a "fair" amount of agreement according to the scale developed

by Cohen (1960). It could be that raters had trouble distinguishing where to place pairs in the five category range. We see some evidence of that when we collapse our five point measure down to three categories (1 = exact or many similar words and phrases in common, 2 =some words and phrases, 3 = a few or no words and phrases) and two categories (1 = exact or many similar words and phrases in common, 2 = some, a few, or no words and phrases in common). There was "moderate" 50% agreement on the three category measure and "substantial" 79% agreement on the two category measure. Each of these was a statistically significant level of agreement.

Figure A5: Human vs. Automated Coding of Bill Pair Similarity



Having achieved some, if not ideal, interrater reliability, we averaged the two rater scores for each bill pair and scaled the five point scale to a 100 point scale, similar to our cosine measure. The human coded bill pair similarity is positively correlated with our cosine similarity measure at r = 0.51, as shown in Figure A5. Although not as strongly related as we would have liked, the human measure and automated measure are on average moving moderately in the expected direction. The moderate human coder evidence and the stronger evidence of cosine similarity tracking strongly with the conceptual definition of reinvention gives us confidence that cosine similarity is a good measure of the concept and that automating the process creates efficiency

and overcomes problems of human coding.

3 Validating Complexity Measure

In this section, we present our results in our tests to verify Flesch Reading Ease (FRE) scores as a measure of complexity. We argue that FRE scores, scaled so more complex bills obtain higher scores, should be a strong and efficient proxy for complexity. The measure is a weighted score of average words per sentence (i.e. long sentences) and average syllables per word (i.e. long words) which should make it more difficult for non-experts to read and understand.

First, we demonstrate how the complexity score applies to real examples from the data. We do this in Figure A6, which shows a section of three bills that establish Safe Haven regulations in their respective states. Each is edited so that section references are skipped and it is formatted as one paragraph; no other changes are made. Each section establishes, more or less, the same policy of no fault for parents of newborns who leave their babies with proper state authorities. But, Arizona (at the 90th percentile in complexity) does so with a long sentence ending in a list of conditions for immunity. Colorado (50th percentile complexity) has a similar long sentence structure but with some plainer language. Maryland (10th percentile complexity), is written directly with even plainer language. Note that the FRE readability scores (ie. the inverse of complexity) are all at the low end of the 121 point scale, high school graduate level reading or higher.

Figure A6: Complexity Example

High Complexity (FRE = 11.68): Arizona Safe Haven Law

A person is not guilty of abuse of a child…solely for leaving an unharmed newborn infant with a safe haven provider. If a parent or agent of a parent voluntarily delivers the parent's newborn infant to a safe haven provider, the safe haven provider shall take custody of the newborn infant if both of the following are true: 1) The parent did not express an intent to return for the newborn infant. 2) The safe haven provider reasonably believes that the child is a newborn infant.

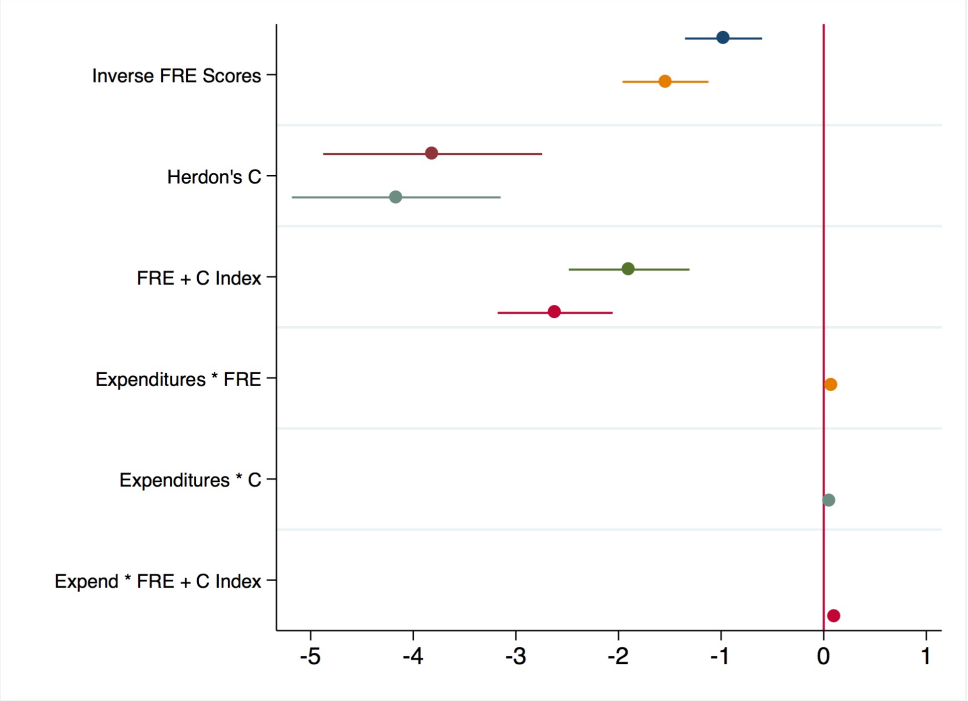
Medium Complexity (FRE = 18.71): Colorado Safe Haven Law

If a parent voluntarily delivers a child to a firefighter…or a hospital staff member who engages in the admission, care, or treatment of patients, when the firefighter is at a fire station or the hospital staff member is at a hospital, the firefighter or hospital staff member shall, without a court order, take temporary physical custody of the child if: 1) The child is seventy-two hours old or younger; and 2) The parent did not express an intent to return for the child.

Low Complexity (FRE = 36.49): Maryland Safe Haven Law

A person who leaves an unharmed newborn with a responsible adult within 10 days after the birth of the newborn, as determined within a reasonable degree of medical certainty, and does not express an intent to return for the newborn shall be immune from civil liability or criminal prosecution for the act.

Figure A7: Coefficient Plot Comparing FRE, C, and FRE + C Index



3.1 Herdan's C

We first check our FRE score measure against another automated measure of complexity, Herdan's C (Herdan 1960; Tweedie and Baayen 1999), which is a logged version of the type-token ratio. C is calculated by taking the log number of types (i.e. unique words) and dividing by the log number of tokes (i.e. total words). The result is the percentage of words in the document that are unique, meaning the reader must track how many different words fit together to create meaning in a document. As shown in Figure A7, we obtain the same substantive results using Herdan's C to the FRE scores, as well as the same results when we combine FRE and C into an additive index. An index that accounts for word rarity, unique words, total words, sentence length, and syllables per word may continue to sharpen our measure of complexity, but likely only slightly and at the cost of efficiency (Benoit, Munger, & Spirling 2019).

3.2 Human Coding

As a further test, we asked human coders to rate the complexity of a random sample of 500 bills. These bills were presented to coders as pairs of bills in the same policy area on the same

survey as the similarity exercise discussed above. We asked "Complexity is how easy or difficult the contents of a bill are to understand. Rate each bill on how difficult or easy it is for you to understand the contents of the bill." Raters were given five options: 1) Extremely easy to understand, 2) somewhat easy to understand, 3) neither easy nor difficult to understand, 4) somewhat difficult to understand, 5) extremely difficult to understand. Two raters coded each bill pair, and we ended up with 476 useable responses in which both raters coded the bill pair without any errors or omissions.

There was very muddled agreement on which bills were complex and which were simple. The raters placed the same bill in the same category just 26% of the time. When we collapse our five point measure down to two categories (1 = extremely, somewhat, or neither easy nor difficult, 2=somewhat or extremely difficult) there was more agreement (77%). This lower than expected agreement also does not correlate much with the automated coding of complexity. Overall, the association between the two is weak and negative (r= -0.03) and raters only agreed with the computer 26% of the time. If we code our bills into two categories based on their automated complexity score (above the median = complex, below = easy) there is still just 55% agreement across the coders and computer.

It appears that these muddled results are due to two reasons 1) the presentation of bills as a pair and 2) trouble keeping consistent coding across policies. The correlation Coder 1's score of bill A and bill B is high (r = 0.67) as is Coder 2's (r = 0.63). Also, it appears that as the correlation between the coder's judgment of the two bills increased, it became more likely that the policy was coded as very similar to the FRE scores or very different from the FRE scores. Lower correlations meant the coders were less certain if the pair was similar or different from the automated measure.

References

Benoit, K., Munger, K., & Spirling, A. (2019). Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, *63*(2), 491-508.

Carley, S., & Miller, C. J. (2012). Regulatory stringency and policy drivers: A reassessment of renewable portfolio standards. *Policy Studies Journal*, *40*(4), 730-56.

Clark, J. (1985). Policy diffusion and program scope: Research directions. *Publius: The Journal of Federalism*, *15*(4), 61-70.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Glick, H. R., & Hays, S. P. (1991). Innovation and reinvention in state policymaking: Theory and the evolution of living will laws. *Journal of Politics*, *53*(3), 835-50.

Hays, S. P. (1996). Patterns of reinvention: The nature of evolution during policy diffusion. *Policy Studies Journal*, *24*(4), 551-66.

Mooney, C. Z., & Lee, M.-H. (1995). Legislative morality in the american states: The case of pre-roe abortion regulation reform. *American Journal of Political Science*, *39*(3), 599-627.