Abstract Readability: Evidence from Top-5 Economics Journals – Supplemental Materials –

Belicia Rodriguez[*] Kim P. Huynh[†] David T. Jacho-Chávez[‡]
Leonardo Sánchez-Aragón[§]

December 26, 2023

1 Data

We proceeded to web scrapped from IDEAS/RePEc and Scopus all articles, notes, erratum, and corrigendum published between January 2000 and December 2019 in the T5 general interest economics journals: *American Economic Review* (AER), *Econometrica* (ECA), the *Journal of Political Economy* (JPE), the *Review of Economic Studies* (RES), and *The Quarterly Journal of Economics* (QJE). This information was then again web scrapped from the journal web pages themselves whenever possible to first cross-verified correct titles, authors' names spelling, missing fields, etc. The final bibliometric dataset contains 5,077 peer-reviewed research articles (hereafter articles) and after excluding non-full length articles, we had a total of 4,988 abstracts to use for our analysis. The following type of papers: short papers, comments, replies, erratum, corrigendum, special issues, and Nobel prize lectures were excluded.

1.1 Authors Gender

We identified a total of 4,884 unique authors who published articles in the T5 this 20-year period. We proceeded to parse their names and extract their first as well as second names (when available). We then used the gender-guesser Python package to classify them as either 'male' or 'female.' There were a total of 4142 authors who were assigned either gender with probability one, while 92 authors were assigned as 'mostly male' and 50 authors were assigned as 'mostly female,' while 600 authors were not possibly assigned by the package. For the latter group we proceeded to do individual internet web pages search and use either photographic publicly available evidence or short biographic information to assign them either gender. We did the same for those assigned as either 'mostly female' and 'mostly male.'

[*]Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: belicia.rodriguez@outlook.com.

[†]Bank of Canada, 234 Wellington Ave., Ottawa ON, K1A 0G9, Canada. E-mail: khuynh@bankofcanada.ca. The views expressed here do not necessarily represent the views of the Bank of Canada.

[‡]Corresponding Author. Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr., Atlanta,GA 30322-2240, USA. E-mail: djachocha@emory.edu

[§]Escuela Superior Politécnica del Litoral, ESPOL, Facultad de Ciencias Sociales y Humanísticas, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador. E-mail: lfsanche@espol.edu.ec.

1.2 Co-Authors Network

Since both IDEAS/RePEc and Scopus provide authors' unique identifiers, these were used to create a co-authorship network among authors who published their work in these T5 journals. An adjacency matrix was then constructed using the `NetworkX` Python package among articles that share at least one co-author in common. This helped us identify a total of 214 disconnected network subgraphs and a total of 542 articles whose authors did not co-author any other T5 papers in these 20 years. We classified all these articles into the 215th group and used it as the base category in our regression analysis.

1.3 JEL Codes

Since only 63% of articles (training data) used in this analysis reported at least one Journal of Economic Literature (JEL) code, we proceeded to imputed the missing JEL codes for the remaining 37% of articles (test data) using a multi-label logistic regression machine learning algorithm from the Scikit-Learn Python package. As predictors we used common words per JEL code as well as the co-authorship network uncovered previously.

Firstly, each abstract was put into a vector representation using the TF-IDF (Term Frequency - Inverse Document Frequency) technique. The TF-IDF value/number increases proportionally with the number of times a word is used in the abstract, but this is balanced with the frequency the same words appear in a collection of papers (shared at least one reported JEL code) - which allow us to identify the most common words per JEL code. The final result is a matrix that is then used as predictors later on. We consider sets of up to 3 words, e.g., "difference in difference."

Secondly, since each article in the training data set has a set of unique co-authors, one can identify the most common JEL code said co-author published in the past and this can be a good predictor of the JEL code we are trying to predict. We achieved an average of 64% prediction accuracy in various training splits before deployment. It means that if the true JEL codes for a given paper are A, B and C, the model predicts JEL codes A and B. There are exactly 3 JEL imputed codes among the 37% articles that did not report any in our final augmented data set.

Please note that this exercise removed JEL codes Y, B, and A since there are only 1, 14, and 26 articles reporting at least one of them.

2 Further Descriptive Statistics

Table 1: Summary Statistics

Variables	Mean	SD	Min	Max	N
Dale-Chall	11.34	0.98	7.37	17.50	4988
American Economic Review	11.41	1.01	7.37	14.30	1500
Econometrica	11.50	0.95	8.23	15.12	892
Journal of Political Economy	11.23	0.97	8.17	14.22	801
Review Economic Studies	11.31	0.93	8.20	15.95	976
The Quarterly Journal of Economics	11.21	1.01	7.64	17.50	819

Note: Descriptive statistics such as sample mean (Mean), standard deviation (SD), minimum (Min), maximum (Max), and sample size (n) for all variables are presented here.

Table 2: Summary Statistics for FK-grade level by JEL code

Variables	Mean	SD	Mean	SD
A	14.91	2.22	14.91	2.22
B	15.13	2.18	15.13	2.18
C	15.76	2.48	15.82	2.56
D	15.41	2.50	15.49	2.58
E	15.49	3.07	15.47	2.92
F	15.43	2.39	15.43	2.47
G	15.33	3.08	15.43	3.05
H	15.41	2.45	15.48	2.49
I	15.41	2.56	15.52	2.73
J	15.17	2.41	15.34	2.57
K	15.26	2.45	15.23	2.64
L	15.29	2.27	15.22	2.33
M	15.41	2.32	15.06	2.26
N	15.45	3.12	15.54	3.03
O	15.51	2.65	15.52	2.74
P	15.87	2.56	15.98	2.52
Q	15.79	2.30	15.75	2.50
R	15.37	2.55	15.40	2.64
Z	15.61	3.09	15.97	3.06

Note: Descriptive statistics such as sample mean (Mean) and standard deviation (SD) are presented here for all JEL codes, except for Y. Only one paper reported that JEL code in the sample period 2000-2019.

Table 3: Summary Statistics for Dale-Chall by JEL code

Variables	Mean	SD	Mean	SD
A	11.10	1.18	11.10	1.18
B	11.15	0.95	11.15	0.95
C	11.44	1.01	11.43	1.00
D	11.36	0.97	11.32	0.97
E	11.39	1.01	11.36	0.99
F	11.39	0.95	11.30	0.97
G	11.44	0.96	11.43	0.96
H	11.34	0.98	11.29	0.95
I	11.33	0.97	11.28	0.97
J	11.30	1.01	11.23	0.99
K	11.50	0.95	11.42	0.96
L	11.51	0.90	11.42	0.91
M	11.31	0.91	11.20	0.89
N	11.20	1.02	11.28	1.04
O	11.44	1.00	11.37	1.00
P	11.40	0.88	11.38	0.87
Q	11.58	0.85	11.50	0.93
R	11.41	0.99	11.37	0.97
Z	11.31	1.04	11.26	1.00

Note: Descriptive statistics such as sample mean (Mean) and standard deviation (SD) are presented here for all JEL codes, except for Y. Only one paper reported that JEL code in the sample period 2000-2019.

3 Further Estimation Results

Table 4: Double-Selection Lasso Linear Estimation Results, JEL codes observed

	(1) log(F-K grade)	(2) log(F-K grade)	(3) log(F-K grade)
log(Number authors)	-0.0046	-0.0051	-0.0047
	(0.0072)	(0.0070)	(0.0072)
log(Number pages)	0.0218**	0.0213**	0.0217**
	(0.0045)	(0.0045)	(0.0045)
Both genders	-0.0124**	-0.0031	0.0130
	(0.0054)	(0.0071)	(0.0100)
Female	-0.0251**		
	(0.0070)		
Share of women		-0.0218**	
		(0.0069)	
Male			0.0256**
			(0.0068)
Observations	3126	3126	3126
Number potential controls	256	256	256
Number controls selected	20	20	16
$\chi^2(4)$	52.270	50.161	51.666

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$

Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test, $\chi^2(4)$, is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.

Table 5: Double-Selection Lasso Linear Estimation Results with Hyperparameters Chosen by Cross-validation & Adaptive Lasso, JEL codes imputed

	(1) log(F-K grade)	(2) log(F-K grade)	(3) log(F-K grade)	(4) log(F-K grade)	(5) log(F-K grade)	(6) log(F-K grade)
log(Number authors)	-0.0053	-0.0056	-0.0054	-0.0056	-0.0053	-0.0056
	(0.0042)	(0.0043)	(0.0042)	(0.0043)	(0.0042)	(0.0043)
log(Number pages)	0.0162**	0.0146**	0.0162**	0.0148**	0.0162**	0.0147**
	(0.0053)	(0.0049)	(0.0053)	(0.0050)	(0.0053)	(0.0050)
Both genders	-0.0070	-0.0070	0.0005	0.0007	0.0128**	0.0134**
	(0.0046)	(0.0049)	(0.0043)	(0.0043)	(0.0062)	(0.0064)
Female	-0.0197**	-0.0198**				
	(0.0058)	(0.0060)				
Share of women			-0.0172**	-0.0177**		
			(0.0058)	(0.0065)		
Male					0.0197**	0.0208**
					(0.0058)	(0.0064)
Observations	4988	4988	4988	4988	4988	4988
Number potential controls	257	257	257	257	257	257
Number controls selected	37	29	37	28	37	29
$\chi^2(4)$	16.657	15.497	13.609	11.591	16.657	14.228
Selection						

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test, $\chi^2(4)$, is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.

Table 6: Double-Selection Lasso Linear, Estimation Results with Hyperparameters Chosen by Cross-validation & Adaptive Lasso, JEL codes observed

	(1) log(F-K grade)	(2) log(F-K grade)	(3) log(F-K grade)	(4) log(F-K grade)	(5) log(F-K grade)	(6) log(F-K grade)
log(Number authors)	-0.0057	-0.0053	-0.0062	-0.0057	-0.0057	-0.0052
	(0.0067)	(0.0071)	(0.0066)	(0.0069)	(0.0067)	(0.0070)
log(Number pages)	0.0192**	0.0197**	0.0192**	0.0197**	0.0192**	0.0198**
	(0.0063)	(0.0069)	(0.0063)	(0.0071)	(0.0063)	(0.0071)
Both genders	-0.0107**	-0.0098*	-0.0006	-0.0005	0.0140	0.0129
	(0.0052)	(0.0050)	(0.0069)	(0.0065)	(0.0110)	(0.0102)
Female	-0.0247**	-0.0225**				
	(0.0079)	(0.0074)				
Share of women			-0.0236**	-0.0216**		
			(0.0074)	(0.0069)		
Male					0.0247**	0.0227**
					(0.0079)	(0.0073)
Observations	3126	3126	3126	3126	3126	3126
Number potential controls	256	256	256	256	256	256
Number controls selected	40	30	40	29	40	29
$\chi^2(4)$	35.370	32.304	35.839	31.542	35.370	31.541
Selection						

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test, $\chi^2(4)$, is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.

Table 7: Cross-Fit Partialing-Out Lasso Linear Estimation Results, JEL codes imputed

	(1) log(F-K grade)	(2) log(F-K grade)	(3) log(F-K grade)
log(Number authors)	-0.0057	-0.0063	-0.0064
	(0.0070)	(0.0070)	(0.0070)
log(Number pages)	0.0171**	0.0171**	0.0174**
	(0.0081)	(0.0081)	(0.0081)
Both genders	-0.0079	0.0010	0.0179*
	(0.0055)	(0.0063)	(0.0102)
Female	-0.0198**		
	(0.0087)		
Share of women		-0.0210**	
		(0.0087)	
Male			0.0262**
			(0.0089)
Observations	4988	4988	4988
Number potential controls	257	257	257
Number controls selected	66	67	67
$\chi^2(4)$	10.452	10.448	13.830

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$

Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test ($\chi^2(4)$) is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.

Table 8: Cross-Fit Partialing-Out Lasso Linear Estimation Results, JEL codes observed

	(1) log(F-K grade)	(2) log(F-K grade)	(3) log(F-K grade)
log(Number authors)	-0.0064	-0.0068	-0.0062
	(0.0068)	(0.0069)	(0.0068)
log(Number pages)	0.0191*	0.0189*	0.0191*
	(0.0101)	(0.0101)	(0.0101)
Both genders	-0.0125*	-0.0017	0.0132
	(0.0066)	(0.0072)	(0.0118)
Female	-0.0274**		
	(0.0118)		
Share of women		-0.0252**	
		(0.0108)	
Male			0.0256**
			(0.0115)
Observations	3126	3126	3126
Number potential controls	256	256	256
Number controls selected	59	60	54
$\chi^2(4)$	11.289	11.488	10.706

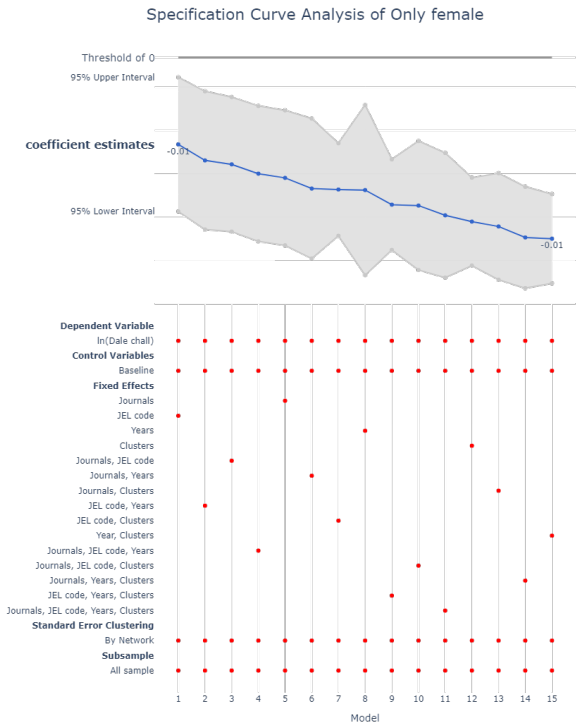
Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

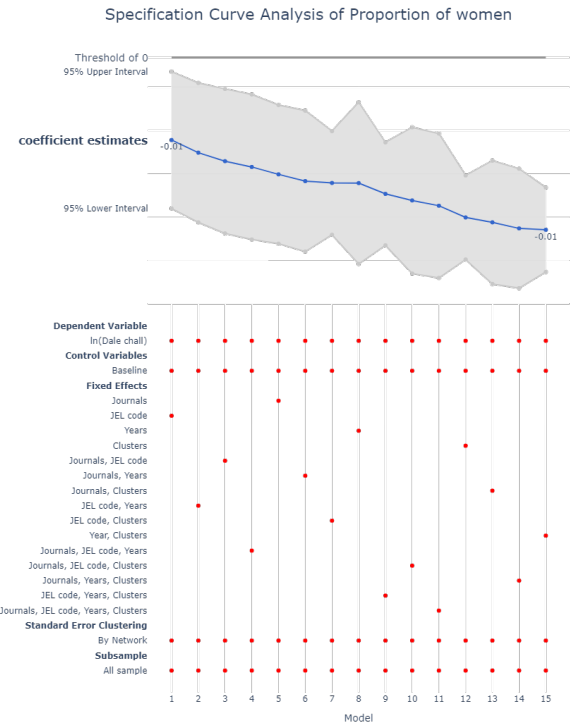
Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test ($\chi^2(4)$) is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.

Figure 1: Specification Curves – Dale-Chall Readability Score

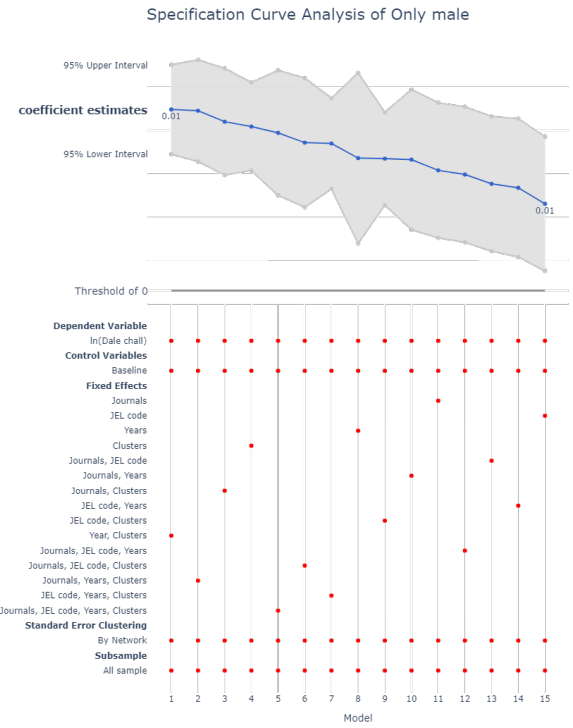
(a) Only Female



(b) Proportion of Women



(c) Only Male



Note: Baseline specification includes an intercept as well as the natural logarithm of the number of authors, word count, and number of pages. It also includes dummies for papers that do not belong to a network subgraph, or do not report any JEL code. Fixed Effects include groups of dummy variables for 4 journals (using the AER as its reference), 19 years (using the year 2000 as its reference), 18 JEL codes (using Microeconomics – D as its reference), and 214 network subgraphs (cluster) membership.

Table 9: Double-Selection Lasso Linear Estimation Results, JEL codes imputed

	(1) log(Dale-Chall)	(2) log(Dale-Chall)	(3) log(Dale-Chall)
log(Number authors)	0.0001	0.0000	0.0001
	(0.0015)	(0.0015)	(0.0015)
log(Number pages)	0.0280**	0.0280**	0.0281**
	(0.0032)	(0.0032)	(0.0032)
Both genders	-0.0069**	-0.0036**	0.0018
	(0.0014)	(0.0017)	(0.0030)
Female	-0.0084**		
	(0.0027)		
Share of women		-0.0076**	
		(0.0025)	
Male			0.0088**
			(0.0026)
Observations	4988	4988	4988
Number potential controls	257	257	257
Number controls selected	21	21	19
$\chi^2(4)$	185.311	143.249	172.661

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test, $\chi^2(4)$, is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.

Table 10: Double-Selection Lasso Linear Estimation Results, JEL codes observed

	(1) log(Dale-Chall)	(2) log(Dale-Chall)	(3) log(Dale-Chall)
log(Number authors)	-0.0005	-0.0005	-0.0004
	(0.0016)	(0.0016)	(0.0016)
log(Number pages)	0.0299**	0.0298**	0.0300**
	(0.0022)	(0.0023)	(0.0022)
Both genders	-0.0052**	-0.0026	0.0020
	(0.0014)	(0.0022)	(0.0045)
Female	-0.0086**		
	(0.0041)		
Share of women		-0.0062	
		(0.0040)	
Male			0.0073*
			(0.0042)
Observations	3126	3126	3126
Number potential controls	256	256	256
Number controls selected	22	22	18
$\chi^2(4)$	209.132	195.321	210.386

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test, $\chi^2(4)$, is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.

Table 11: Double-Selection Lasso Linear Estimation Results with Hyperparameters Chosen by Cross-validation & Adaptive Lasso, JEL codes imputed

	(1) log(Dale-Chall)	(2) log(Dale-Chall)	(3) log(Dale-Chall)	(4) log(Dale-Chall)	(5) log(Dale-Chall)	(6) log(Dale-Chall)
log(Number authors)	-0.0017	-0.0020	-0.0017	-0.0021	-0.0017	-0.0020
	(0.0019)	(0.0018)	(0.0019)	(0.0018)	(0.0019)	(0.0018)
log(Number pages)	0.0267**	0.0262**	0.0267**	0.0260**	0.0267**	0.0260**
	(0.0039)	(0.0038)	(0.0039)	(0.0037)	(0.0039)	(0.0037)
Both genders	-0.0064**	-0.0070**	-0.0030**	-0.0034**	0.0025	0.0024
	(0.0014)	(0.0015)	(0.0014)	(0.0014)	(0.0024)	(0.0025)
Female	-0.0090**	-0.0098**				
	(0.0027)	(0.0027)				
Share of women			-0.0080**	-0.0084**		
			(0.0027)	(0.0027)		
Male					0.0090**	0.0093**
					(0.0027)	(0.0027)
Observations	4988	4988	4988	4988	4988	4988
Number potential controls	257	257	257	257	257	257
Number controls selected	37	29	37	28	37	29
$\chi^2(4)$	80.768	74.107	83.054	76.220	80.768	75.950
Selection						

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test, $\chi^2(4)$, is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.

Table 12: Double-Selection Lasso Linear Estimation Results with Hyperparameters Chosen by Cross-validation & Adaptive Lasso, JEL codes observed

	(1) log(Dale-Chall)	(2) log(Dale-Chall)	(3) log(Dale-Chall)	(4) log(Dale-Chall)	(5) log(Dale-Chall)	(6) log(Dale-Chall)
log(Number authors)	-0.0020	-0.0019	-0.0020	-0.0021	-0.0020	-0.0021
	(0.0018)	(0.0017)	(0.0019)	(0.0018)	(0.0018)	(0.0017)
log(Number pages)	0.0270**	0.0273**	0.0270**	0.0269**	0.0270**	0.0269**
	(0.0041)	(0.0045)	(0.0041)	(0.0043)	(0.0041)	(0.0043)
Both genders	-0.0054**	-0.0055**	-0.0020	-0.0019	0.0040	0.0042
	(0.0015)	(0.0015)	(0.0019)	(0.0019)	(0.0033)	(0.0037)
Female	-0.0095**	-0.0101**				
	(0.0031)	(0.0033)				
Share of women			-0.0079**	-0.0083**		
			(0.0030)	(0.0032)		
Male					0.0095**	0.0097**
					(0.0031)	(0.0034)
Observations	3126	3126	3126	3126	3126	3126
Number potential controls	256	256	256	256	256	256
Number controls selected	40	30	40	29	40	29
$\chi^2(4)$	72.940	69.792	68.237	61.225	72.940	66.921
Selection						

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test, $\chi^2(4)$, is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.

Table 13: Cross-Fit Partialing-Out Lasso Linear Estimation Results, JEL codes imputed

	(1) log(Dale-Chall)	(2) log(Dale-Chall)	(3) log(Dale-Chall)
log(Number authors)	0.0000	-0.0003	-0.0001
	(0.0030)	(0.0030)	(0.0030)
log(Number pages)	0.0279**	0.0279**	0.0280**
	(0.0045)	(0.0045)	(0.0045)
Both genders	-0.0071**	-0.0031	0.0029
	(0.0030)	(0.0031)	(0.0053)
Female	-0.0087		
	(0.0054)		
Share of women		-0.0093*	
		(0.0051)	
Male			0.0101*
			(0.0052)
Observations	4988	4988	4988
Number potential controls	257	257	257
Number controls selected	67	69	70
$\chi^2(4)$	55.253	57.882	55.835

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test ($\chi^2(4)$) is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.

Table 14: Cross-Fit Partialing-Out Lasso Linear Estimation Results, JEL codes observed

	(1) log(Dale-Chall)	(2) log(Dale-Chall)	(3) log(Dale-Chall)
log(Number authors)	-0.0025	-0.0024	-0.0023
	(0.0046)	(0.0046)	(0.0046)
log(Number pages)	0.0293**	0.0293**	0.0293**
	(0.0066)	(0.0066)	(0.0066)
Both genders	-0.0058	-0.0033	0.0004
	(0.0043)	(0.0058)	(0.0094)
Female	-0.0077		
	(0.0076)		
Share of women		-0.0057	
		(0.0069)	
Male			0.0061
			(0.0075)
Observations	3126	3126	3126
Number potential controls	256	256	256
Number controls selected	59	60	54
$\chi^2(4)$	28.098	29.147	26.640

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$

Note: Clustered standard errors at the network disconnected subgraph level are in parentheses. * p-value < 0.10, ** p-value < 0.05. The Chi-squared test ($\chi^2(4)$) is a Wald test of the coefficients of the 4 variables of interest jointly equal to zero in each specification.