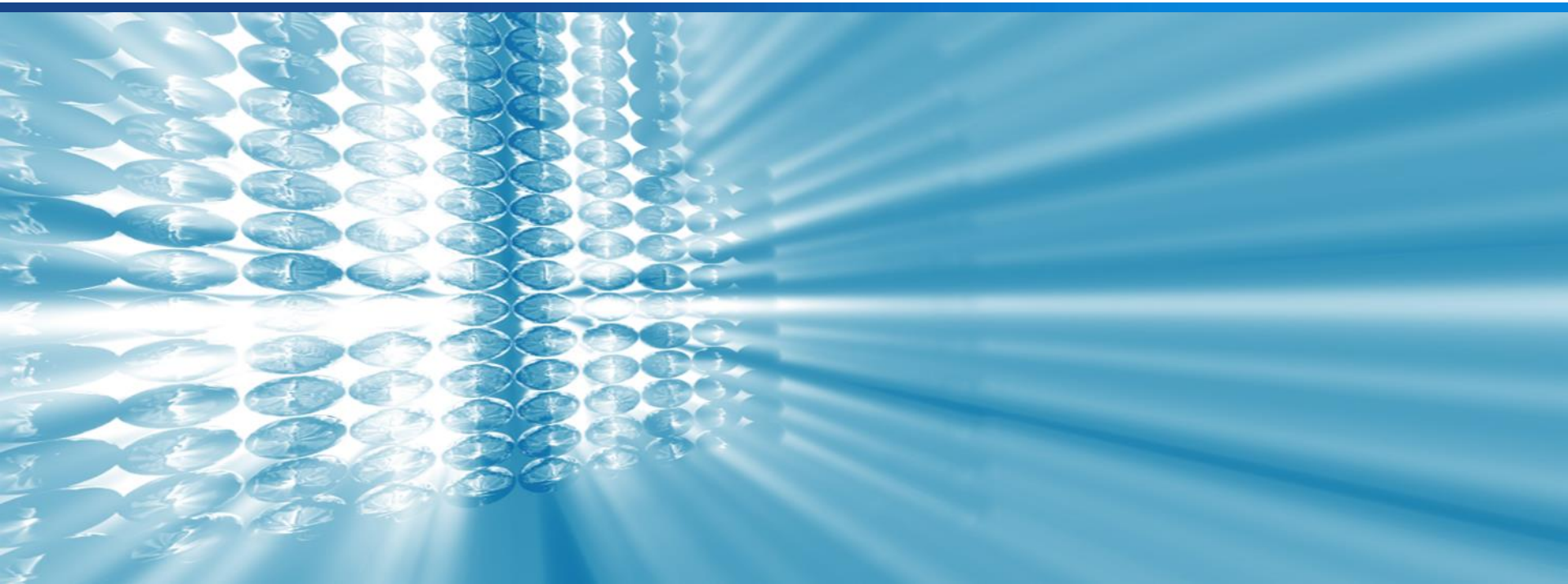
Science-Metrix

Influence of OA, Gender, Co-authorship on Citation

Science & Technology Indicators 2018 (Leiden)

September 12–14, 2018



Game plan

- Study context
- Study design
- Analysis
  - Objective 1: replicate previous findings in same 2 subfields
  - Objective 2: expand scope of analysis
  - Objective 3: drill down into local contexts
- Conclusions

Game plan

- Study context
- Study design
- Analysis
  - Objective 1: replicate previous findings in same 2 subfields
  - Objective 2: expand scope of analysis
  - Objective 3: drill down into local contexts
- Conclusions

Study context

- Citations used to evaluate (academic) impact of research.
- Valuable to understand citation determinants, for indicator normalization and to guide research policy.
- 3 parameters that are known to influence citation rates:
  - International collaboration
  - Open access
  - Gender composition of research team
- These 3 parameters are also inter-related.
- So what's actually having an effect?

Study context

- Last year at STI in Paris, we presented preliminary results: analyses were limited to 1 year, 2 subfields, small data set.
- Major point of discussion at STI 2017 and ISSI 2017 was about replicability and robustness of findings.
- So for STI 2018, we put our previous results to the test!
- Replication is conceptual rather than exact; convergence would lend additional credence to the findings.

Game plan

- Study context
- Study design
- Analysis
  - Objective 1: replicate previous findings in same 2 subfields
  - Objective 2: expand scope of analysis
  - Objective 3: drill down into local contexts
- Conclusions

Study design

- Data sources:
  - Web of Science, produced by Clarivate (last time was Scopus)
  - 1findr, produced by 1science
  - NamSor API for data enrichment (gender tagging of author names)
- Coverage & filters:
  - 2008–2012 (last time was just 2012)
  - All subfields, except Arts & Humanities (last time just 2 subfields)
  - Journals not on 1science “whitelist” excluded
  - All authors on a paper must be tagged by NamSor (not 100% conf.)

Study design

- Parametrization
  - OA: gold, green, gold + green, unknown (last time OA was binary)
  - Gender bins: <20% women, 20%–40%, 40%–60%, 60%–80%, >80% (last time was binary + scalar)
  - # authors, institutions: scalar
  - International collaboration: binary
- Modelling approaches:
  - Negative binomial (treats citations as count variable)
  - Robust (resilient against non-normal distributions in inputs, outputs)
  - Zero-inflated negative binomial (models 0's separately from rest)

Game plan

- Study context
- Study design
- Analysis
  - Objective 1: replicate previous findings in same 2 subfields
  - Objective 2: expand scope of analysis
  - Objective 3: drill down into local contexts
- Conclusions

Analysis 1: replicate previous findings

Previous findings—

- OA:
  - Positive, significant, meaningful effect on citation scores
- International collaboration:
  - Positive, meaningful effect on citation scores
  - Smaller than OA's effect
  - Significant in cardio system & hematology, not in dev. biology.
- Gender composition:
  - Mixed-gender teams had strongest effect, but optimal mix not clear.
  - Significant in dev. biology, not in cardio system & hematology

Analysis 1: replicate previous findings

This time around—

- OA:
  - Positive, significant, meaningful effect on citation scores
  - Again in both subfields
  - Again strongest effect detected
  - Consistent across all three models
- International collaboration:
  - Positive, significant, meaningful effect
  - In both subfields; smaller magnitude where previously not significant
  - Consistent across all three models
- Gender composition:
  - Dev. bio: negative, significant, meaningful effect across models
  - Cardio: positive, significant (Robust model only), meaningful effect

Analysis 1: replicate previous findings

Developmental biology

	Variable	Type	Predicted benefit: Negative binomial	Predicted benefit: Robust regression	Predicted benefit: Zero-inflated	Coefficients: Negative binomial	Coefficients: Robust regression	Coefficients: Zero-inflated	Stat. significance: Negative binomial	Stat. significance: Robust regression	Stat. significance: Zero-inflated
Open access	Gold	Binary	38%	66%	38%	0.323	0.505	0.321	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Green	Binary	48%	45%	47%	0.390	0.374	0.388	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Green + Gold	Binary	60%	71%	59%	0.497	0.539	0.465	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Unknown	Binary	46%	35%	46%	0.378	0.302	0.377	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
% women in team	<20%	Binary	0%	0%	0%	n/a	n/a	n/a	n/a	n/a	n/a
	20%–40%	Binary	-13%	-7%	-13%	-0.145	-0.070	-0.145	p < 2.00e-16	p = 1.61e-07	p < 2.00e-16
	40%–60%	Binary	-20%	-12%	-20%	-0.218	-0.128	-0.218	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	60%–80%	Binary	-33%	-20%	-33%	-0.400	-0.226	-0.400	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	>80%	Binary	-31%	-22%	-31%	-0.377	-0.254	-0.372	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
Collab.	Number of authors	Scalar	4%	6%	4%	0.042	0.054	0.042	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Number of addresses	Scalar									
	International collab.	Binary	15%	16%	15%	0.140	0.151	0.140	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16

Cardiovascular system & hematology

	Variable	Type	Predicted benefit: Negative binomial	Predicted benefit: Robust regression	Predicted benefit: Zero-inflated	Coefficients: Negative binomial	Coefficients: Robust regression	Coefficients: Zero-inflated	Stat. significance: Negative binomial	Stat. significance: Robust regression	Stat. significance: Zero-inflated
Open access	Gold	Binary	113%	124%	111%	0.757	0.805	0.748	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Green	Binary	57%	74%	55%	0.454	0.556	0.438	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Green + Gold	Binary	145%	147%	144%	0.896	0.906	0.894	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Unknown	Binary	44%	50%	43%	0.366	0.404	0.355	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
% women in team	<20%	Binary	0%	0%	0%	n/a	n/a	n/a	n/a	n/a	n/a
	20%–40%	Binary	2%	9%	2%	0.021	0.085	0.017	p = 4.59e-02	p = 2.26e-12	p = 1.08e-01
	40%–60%	Binary	3%	13%	2%	0.033	0.126	0.024	p = 8.77e-03	p < 2.00e-16	p = 6.16e-02
	60%–80%	Binary	1%	14%	0%	0.013	0.131	-0.001	p = 4.92e-01	p = 1.85e-09	p = 9.41e-01
	>80%	Binary	4%	18%	2%	0.040	0.163	0.021	p = 1.64e-01	p = 1.77e-06	p = 4.64e-01
Collab.	Number of authors	Scalar	4%	6%	4%	0.038	0.060	0.036	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Number of addresses	Scalar	5%	5%	5%	0.053	0.051	0.051	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	International collab.	Binary	27%	39%	26%	0.242	0.330	0.235	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16

Game plan

- Study context
- Study design
- Analysis
  - Objective 1: replicate previous findings in same 2 subfields
  - Objective 2: expand scope of analysis
  - Objective 3: drill down into local contexts
- Conclusions

Analysis 2: expand scope

- OA:
  - Positive, significant, meaningful effect on citation scores
  - Again strongest effect detected
  - Much stronger for green OA than gold OA
  - Consistent across all three models
- International collaboration:
  - Positive, significant meaningful effect
  - Consistent across all three models
- Gender composition:
  - Citation penalty as share of women increases
  - Smaller magnitude than either international collab. or OA effects
  - 2 of 3 models agree (Robust logs the “minority report”)

Analysis 2: expand scope

All fields, worldwide

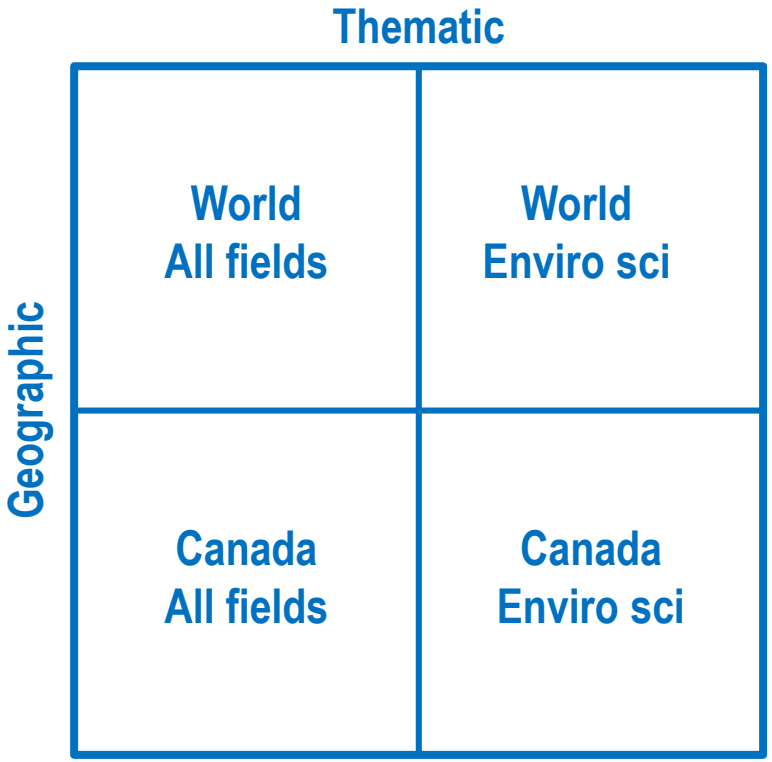
	Variable	Type	Predicted benefit			Coefficients			Stat. significance		
			Negative binomial	Robust regression	Zero-inflated	Negative binomial	Robust regression	Zero-inflated	Negative binomial	Robust regression	Zero-inflated
Open access	Gold	Binary	15%	11%	15%	0.142	0.106	0.142	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Green	Binary	55%	65%	54%	0.436	0.499	0.429	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Green + Gold	Binary	44%	45%	44%	0.367	0.372	0.362	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Unknown	Binary	45%	47%	45%	0.374	0.385	0.369	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
% women in team	<20%	Binary	0%	0%	0%	n/a	n/a	n/a	n/a	n/a	n/a
	20%–40%	Binary	-3%	3%	-4%	-0.031	0.033	-0.036	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	40%–60%	Binary	-6%	1%	-6%	-0.061	0.006	-0.064	p < 2.00e-16	p = 6.60e-03	p < 2.00e-16
	60%–80%	Binary	-10%	-1%	-11%	-0.107	-0.010	-0.113	p < 2.00e-16	p = 1.68e-04	p < 2.00e-16
	>80%	Binary	-12%	-6%	-12%	-0.128	-0.057	-0.123	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
Collab.	Number of authors	Scalar	3%	4%	2%	0.025	0.041	0.023	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Number of addresses	Scalar	3%	3%	4%	0.034	0.025	0.035	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	International collab.	Binary	18%	23%	17%	0.162	0.208	0.161	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16

Game plan

- Study context
- Study design
- Analysis
  - Objective 1: replicate previous findings in same 2 subfields
  - Objective 2: expand scope of analysis
  - Objective 3: drill down into local contexts
- Conclusions

Analysis 3: drill down

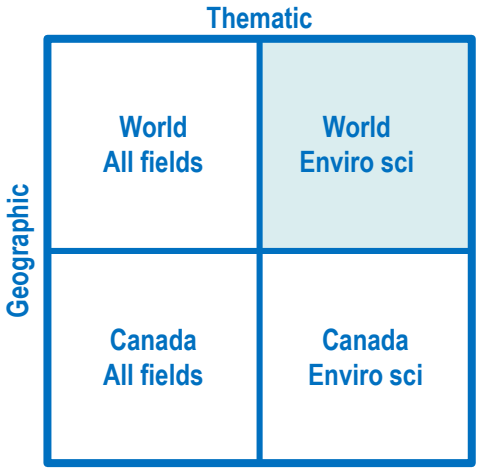
- Looked at 1 thematic subset, 1 geographical subset, crossover
- Compared results to global findings



Analysis 3: drill down

Environmental science, worldwide—

- OA:
  - Citation advantage smaller in enviro sci than elsewhere
  - Gold OA even has citation penalty here!
  - Consistent across models
- International collaboration:
  - Smaller citation advantage than elsewhere
  - Consistent across models
- Gender composition:
  - More women, lower citation score
  - Consistent across models



Analysis 3: drill down

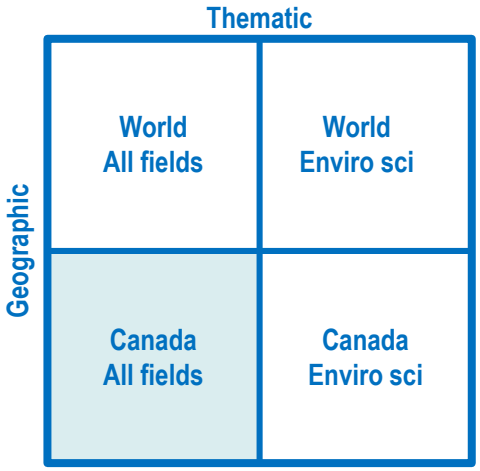
Environmental science, worldwide

	Variable	Type	Predicted benefit			Coefficients			Stat. significance		
			Negative binomial	Robust regression	Zero-inflated	Negative binomial	Robust regression	Zero-inflated	Negative binomial	Robust regression	Zero-inflated
Open access	Gold	Binary	-12%	-30%	-11%	-0.128	-0.354	-0.119	p = 3.21e-16	p < 2.00e-16	p = 2.74e-13
	Green	Binary	36%	38%	36%	0.308	0.319	0.306	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Green + Gold	Binary	21%	4%	21%	0.190	0.043	0.188	p < 2.00e-16	p = 4.45e-02	p < 2.00e-16
	Unknown	Binary	44%	41%	44%	0.366	0.347	0.364	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
% women in team	<20%	Binary	0%	0%	0%	n/a	n/a	n/a	n/a	n/a	n/a
	20%–40%	Binary	-3%	0%	-3%	-0.032	0.000	-0.033	p = 5.36e-05	p = 9.73e-01	p = 2.81e-05
	40%–60%	Binary	-5%	-3%	-5%	-0.050	-0.030	-0.049	p = 2.34e-08	p = 2.26e-03	p = 5.64e-08
	60%–80%	Binary	-8%	-5%	-8%	-0.084	-0.055	-0.085	p = 6.11e-10	p = 1.70e-04	p = 5.84e-10
	>80%	Binary	-13%	-11%	-13%	-0.142	-0.122	-0.139	p = 9.25e-16	p = 2.00e-09	p = 7.77e-15
Collab.	Number of authors	Scalar	5%	6%	4%	0.045	0.054	0.044	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Number of addresses	Scalar	2%	2%	2%	0.024	0.019	0.024	p = 6.48e-16	p = 7.2.e-09	p < 2.00e-16
	International collab.	Binary	11%	18%	10%	0.100	0.165	0.098	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16

Analysis 3: drill down

Canadian research—

- OA:
  - Clear OA citation advantage.
  - Gold is same magnitude as elsewhere; green smaller (still > gold)
  - Consistent across models
- International collaboration:
  - Same citation advantage as elsewhere
  - Consistent across models
- Gender composition:
  - All-women teams fared better here than elsewhere
  - Mixed-gender teams fared marginally worse
  - 2 of 3 models agree (Robust is dissenter again)



Analysis 3: drill down

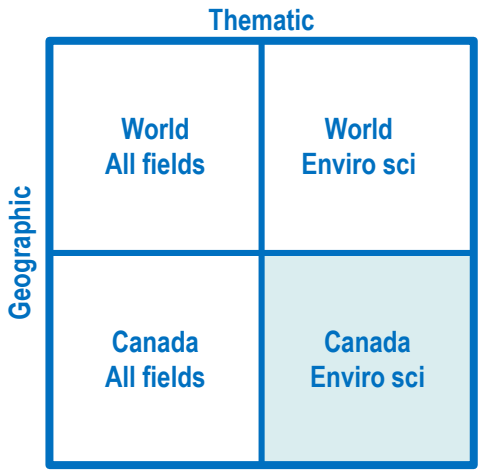
All fields, Canada

	Variable	Type	Predicted benefit: Negative binomial	Predicted benefit: Robust regression	Predicted benefit: Zero-inflated	Coefficients: Negative binomial	Coefficients: Robust regression	Coefficients: Zero-inflated	Stat. significance: Negative binomial	Stat. significance: Robust regression	Stat. significance: Zero-inflated
Open access	Gold	Binary	15%	12%	15%	0.139	0.117	0.140	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Green	Binary	40%	40%	40%	0.337	0.338	0.334	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Green + Gold	Binary	31%	28%	30%	0.267	0.247	0.265	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Unknown	Binary	42%	40%	42%	0.353	0.333	0.351	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
% women in team	<20%	Binary	0%	0%	0%	n/a	n/a	n/a	n/a	n/a	n/a
	20%–40%	Binary	-5%	0%	-5%	-0.051	-0.003	-0.053	p = 2.58e-11	p = 7.26e-01	p = 3.86e-12
	40%–60%	Binary	-8%	-3%	-8%	-0.079	-0.034	-0.081	p < 2.00e-16	p = 8.07e-05	p < 2.00e-16
	60%–80%	Binary	-11%	-3%	-11%	-0.115	-0.031	-0.118	p < 2.00e-16	p = 4.91e-03	p < 2.00e-16
	>80%	Binary	-9%	-2%	-8%	-0.089	-0.022	-0.088	p = 9.29e-14	p = 9.24e-02	p = 1.50e-13
Collab.	Number of authors	Scalar	4%	5%	4%	0.040	0.050	0.040	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	Number of addresses	Scalar	4%	2%	4%	0.035	0.021	0.036	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16
	International collab.	Binary	17%	16%	17%	0.155	0.149	0.155	p < 2.00e-16	p < 2.00e-16	p < 2.00e-16

Analysis 3: drill down

Canadian environmental science—

- OA:
  - Gold OA penalty even more intense here than enviro sci globally. And Canada across fields enjoyed a strong gold OA advantage!
  - Green OA advantage about 15 points lower than Canadian level or enviro sci level worldwide.
  - Consistent across models
- International collaboration:
  - Larger advantage than Canada (+5) or env sci (+10)
  - Consistent across models
- Gender composition:
  - Teams with more women fared best.
  - Consistent across models, but stat. sig. degraded
  - Few observations of women-led teams—worth replication, explanation!



Analysis 3: drill down

Environmental science, Canada

	Variable	Type	Predicted benefit: Negative binomial	Predicted benefit: Robust regression	Predicted benefit: Zero-inflated	Coefficients: Negative binomial	Coefficients: Robust regression	Coefficients: Zero-inflated	Stat. significance: Negative binomial	Stat. significance: Robust regression	Stat. significance: Zero-inflated
Open access	Gold	Binary	-33%	-40%	-34%	-0.406	-0.510	-0.410	p = 1.21e-06	p = 5.94e-06	p = 7.71e-07
	Green	Binary	23%	22%	22%	0.205	0.200	0.201	p = 3.38e-13	p = 1.02e-11	p = 8.46e-13
	Green + Gold	Binary	11%	22%	16%	0.106	0.199	0.145	p = 1.90e-1	p = 1.95e-02	p = 8.07e-02
	Unknown	Binary	24%	17%	24%	0.216	0.161	0.212	p = 2.84e-08	p = 2.01e-04	p = 4.70e-08
% women in team	<20%	Binary	0%	0%	0%	n/a	n/a	n/a	n/a	n/a	n/a
	20%–40%	Binary	-9%	-5%	-10%	-0.098	-0.051	-0.102	p = 1.47e-03	p = 1.11e-01	p = 9.20e-04
	40%–60%	Binary	-9%	-5%	-10%	-0.096	-0.051	-0.100	p = 7.47e-03	p = 1.83e-01	p = 4.74e-03
	60%–80%	Binary	10%	12%	9%	0.091	0.116	0.086	p = 9.61e-02	p = 4.48e-02	p = 1.15e-01
	>80%	Binary	4%	2%	4%	0.043	0.024	0.036	p = 6.12e-01	p = 8.10e-01	p = 6.68e-01
Collab.	Number of authors	Scalar	5%	7%	5%	0.052	0.067	0.050	p = 1.58e-08	p = 5.25e-12	p = 6.70e-08
	Number of addresses	Scalar	4%	2%	4%	0.036	0.019	0.037	p = 4.56e-03	p = 1.54e-01	p = 3.61e-03
	International collab.	Binary	22%	23%	22%	0.197	0.211	0.200	p = 1.15e-11	p = 1.67e-11	p = 5.27e-12

Game plan

- Study context
- Study design
- Analysis
  - Objective 1: replicate previous findings in same 2 subfields
  - Objective 2: expand scope of analysis
  - Objective 3: drill down into local contexts
- Conclusions

Conclusions

- Previous findings pretty much all corroborated, even with different data source, parametrization, modelling.
- Wider analyses discovered similar patterns elsewhere.
- Takeaways about citation determinants:
  - OA citation advantage strong, esp. for green OA
  - International collaboration advantage clear also, smaller than OA
  - As share of women increases, citation numbers decrease.
  - Drilling down into subsets offers valuable context for interpretation.
- Takeaways about bibliometric modelling:
  - Important to consider magnitudes of effect, not just significance!
  - Convergence across modelling approaches helps to assess reliability/spuriousness of various findings.
  - Valuable to try out new data sources

Conclusions

- Reflections on research policy:
  - Modelling such as this can be used to establish more flexible benchmarks against which to measure performance.
  - Should we be holding different people to different standards?
  - Also raises question about what is a “legitimate” research strategy as opposed to “gaming” the system.
  - What perverse incentives might flexible systems introduce?
  - These issues are raised—not resolved!—by the present study.

Thank you!

CONTACT

Brooke Struck
Senior policy officer
brooke.struck@science-metrix.com

Matthew Durning
Research analyst

Guillaume Roberge
Senior analyst

David Campbell
Chief scientist

Science-Metrix
Montréal

WEBSITE
www.science-metrix.com

PHONE
1.514.495.6505
1.800.994.4761