The Effect on Citation Inequality of Differences in Citation Practices across Scientific Fields

Juan A. Crespo[1], Yunrong Li[2], Javier Ruiz-Castillo[2]

[2]Universidad Carlos III de Madrid, Spain

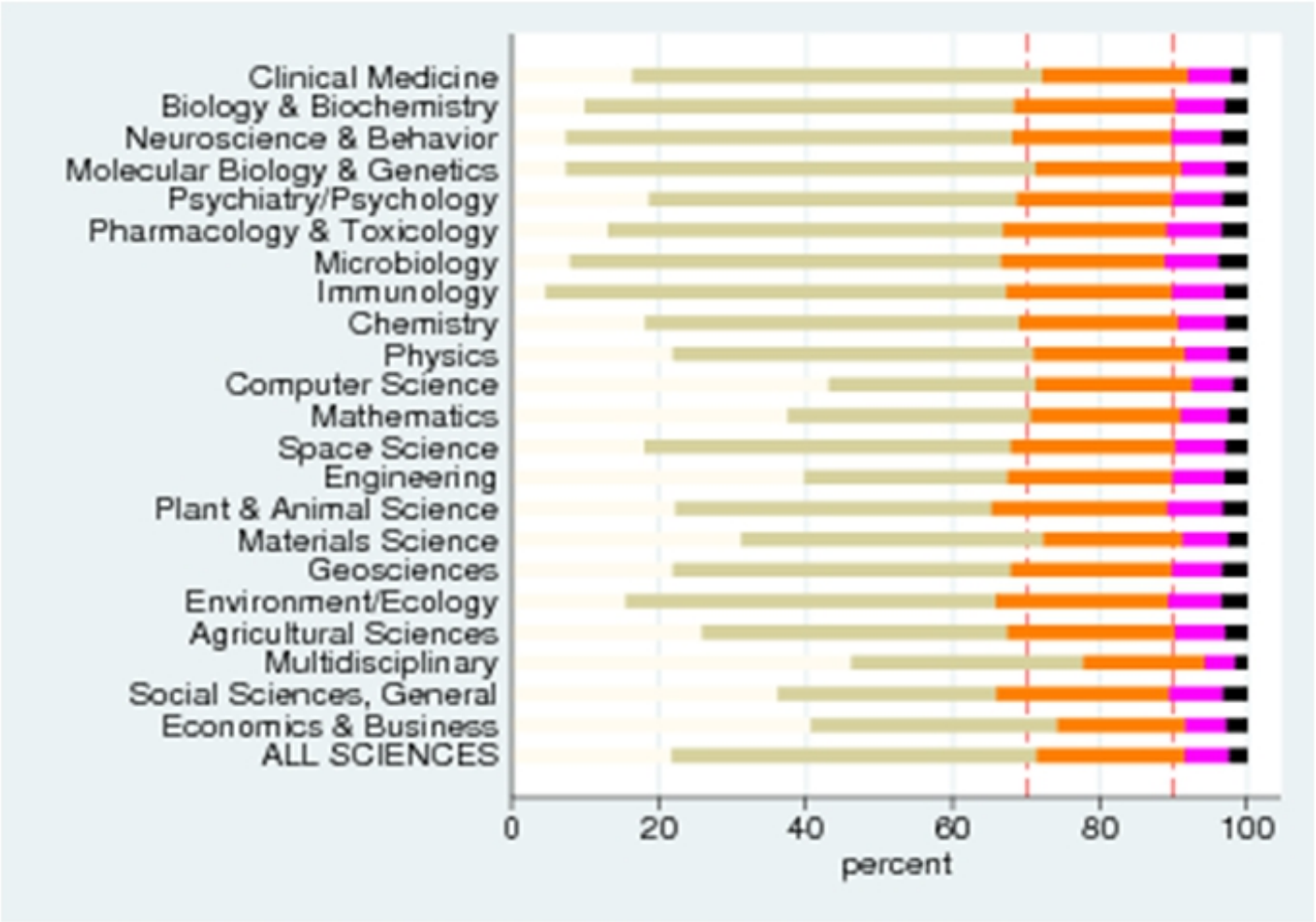
October 10, 2013

Outline

Motivation

- How to assess the scientific influence of a research paper?
  - Citation impact: the number of citations received by the paper within a certain period of time after its publication.
- A Scientific Field: a collection of papers published in a set of closely related professional journals.
- Empirical Regularity: highly skewed citation distributions (Albarrán,2011), citation inequality is very large within a field as well as in all-fields case.

Skewness of Citation Distribution



Motivation

- The large citation inequality may be due to different papers have different scientific influence, or papers belong to different fields.
- The field dependence of citation impacts:
  - Size of the field: average number of papers per author in a given period of time.
  - Average number of references per paper.
  - The speed at which the citation process evolves.
- To introduce a simple model to see how important is the field dependence of citation impacts on citation inequality.

Motivation

- The large citation inequality may be due to different papers have different scientific influence, or papers belong to different fields.
- The field dependence of citation impacts:
  - Size of the field: average number of papers per author in a given period of time.
  - Average number of references per paper.
  - The speed at which the citation process evolves.
- To introduce a simple model to see how important is the field dependence of citation impacts on citation inequality.

Motivation

- The large citation inequality may be due to different papers have different scientific influence, or papers belong to different fields.
- The field dependence of citation impacts:
  - Size of the field: average number of papers per author in a given period of time.
  - Average number of references per paper.
  - The speed at which the citation process evolves.
- To introduce a simple model to see how important is the field dependence of citation impacts on citation inequality.

Statistics of Field Citation Distribution

Table A. Number of Articles and Mean Citation Rates by Field

	Number of Articles	%	Mean Citation	Standard Deviation	CV
A. LIFE SCIENCES	**1,806,398**	**40.4**			
1. Biology & Biochemistry	275,568	6.2	12.6	20.1	1.6
2. Clinical Medicine	947,261	21.2	9.7	21.6	2.2
3. Immunology	60,875	1.4	16.0	23.0	1.4
4. Microbiology	73,039	1.6	11.4	13.9	1.2
5. Molecular Biology & Genetics	122,233	2.7	20.4	32.7	1.6
6. Neuroscience & Behav. Science	140,686	3.2	13.7	18.2	1.3
7. Pharmacology & Toxicology	76,728	1.7	8.0	11.0	1.4
8. Psychiatry & Psychology	110,008	2.5	7.0	11.3	1.6
B. PHYSICAL SCIENCES	**1,282,919**	**28.7**			
9. Chemistry	550,147	12.3	7.6	14.2	1.9
10. Computer Science	98,727	2.2	3.0	13.8	4.7
11. Mathematics	117,496	2.6	2.5	5.2	2.1
12. Physics	456,144	10.2	6.9	14.9	2.2
13. Space Science	60,405	1.4	11.0	20.5	1.9
C. OTHER NATURAL SCIENCES	**1,150,428**	**25.7**			
14. Agricultural Sciences	82,837	1.9	4.9	7.2	1.5
15. Engineering	356,269	8.0	3.2	5.8	1.8
16. Environment & Ecology	109,826	2.5	7.1	10.3	1.4
17. Geoscience	120,059	2.7	6.7	10.0	1.5
18. Materials Science	199,364	4.5	4.5	8.9	2.0
19. Multidisciplinary	20,672	0.5	3.2	7.0	2.2
20. Plant & Animal Science	261,401	5.8	5.1	8.0	1.6
D. SOCIAL SCIENCES	**232,587**	**5.2**			
21. Economics & Business	63,380	1.4	4.0	7.1	1.8
22. Social Sciences, General	169,207	3.8	3.3	5.7	1.7
ALL FIELDS	**4,472,332**	**100**	**7.8**	**13.2**	**1.8**

Roemer´s Income Inequality Model

- Roemer´s (1998) model:
  - Income of individuals depends on efforts and circumstances (e.g. parents´ wealth, education.), partition the population by "type",

$$Income_{ti} = (type_t, effort_{ti}).$$

  - Effort distribution within a type is a characteristic of the type.
  - A1: Within a type, individuals at the same quantile of effort distribution implement the same "degree" of effort.
  - A2: Within a type, income is monotonic in effort. Quantiles of effort distribution correspond to quantiles of income distribution.
  - Holding constant the degree of effort/income, income inequality across types is due to circumstances, "inequality of opportunity".

Analogy of Our Model with Roemer´s

- Individuals$\Longrightarrow$Articles
- Income$\Longrightarrow$Citation impact
- A income distribution$\Longrightarrow$A citation distribution
- Circumstances/Types$\Longrightarrow$Fields
- Effort$\Longrightarrow$Scientific influence

Assumptions

- $Citation_{fi} = (field_f,\ Scientific\ Influence_{fi}),\ f = 1, ..., F;\ i = 1, ..., N.$
- A1: Within a field, articles at the same quantile of the scientific influence distribution reflect the same "degree" of scientific influence.
- A2: $Citation_{fi}$ is monotonic in $scientific\ influence_{fi}$.
  - Quantiles of scientific influence distribution correspond the quantiles of citation distribution.

Assumptions

- $Citation_{fi} = (field_f,\ Scientific\ Influence_{fi}),\ f = 1, ..., F;\ i = 1, ..., N.$
- A1: Within a field, articles at the same quantile of the scientific influence distribution reflect the same "degree" of scientific influence.
- A2: $Citation_{fi}$ is monotonic in $scientific\ influence_{fi}$.
  - Quantiles of scientific influence distribution correspond the quantiles of citation distribution.

Assumptions

- $Citation_{fi} = (field_f,\ Scientific\ Influence_{fi}),\ f = 1, ..., F;\ i = 1, ..., N.$
- A1: Within a field, articles at the same quantile of the scientific influence distribution reflect the same "degree" of scientific influence.
- A2: $Citation_{fi}$ is monotonic in $scientific\ influence_{fi}$.
  - Quantiles of scientific influence distribution correspond the quantiles of citation distribution.

Double Partition

- Partition the all-fields citation distribution by fields f and quantiles π.

$$
\begin{array}{c}
\textit{Field } 1 \\
\left\{ \begin{array}{ll} 1st & \\ 2nd & \\ \vdots & \textit{quantiles} \\ \Pi th & \end{array} \right. \\
\vdots \\
\textit{Field } F \\
\left\{ \begin{array}{ll} 1st & \\ 2nd & \\ \vdots & \textit{quantiles} \\ \Pi th & \end{array} \right.
\end{array}
$$

Double Partition

- "sort", citations are in asending order within each field.
- "_pctile", create 1000 quantiles within each field.
- "merge" to merge all fields together.
- All-fields citation distribution is a matrix of cells, (f, π).

Decomposable Inequality Index

- Generalized Entropy (GE) family of inequality indices, the Theil index.
  - $I_1(C) = \frac{1}{N}\sum_I \frac{c_I}{\mu} log \frac{c_I}{\mu}$, citation inequality of all-fields case, μ is the mean citation of all articles.
  - For articles with 0 ciation, we follow the convention $0 * log 0 = 0$.
  - $I_1(C)$ is decomposable, $I_1(C) = W + S + IDCP$.

Decomposable Inequality Index

- Generalized Entropy (GE) family of inequality indices, the Theil index.
  - $I_1(C) = \frac{1}{N}\sum_I \frac{c_I}{\mu} log \frac{c_I}{\mu}$, citation inequality of all-fields case, μ is the mean citation of all articles.
  - For articles with 0 ciation, we follow the convention $0 * log 0 = 0$.
  - $I_1(C)$ is decomposable, $I_1(C) = W + S + IDCP$.

Decomposable Inequality Index

- Generalized Entropy (GE) family of inequality indices, the Theil index.
  - $I_1(C) = \frac{1}{N}\sum_I \frac{c_I}{\mu} log \frac{c_I}{\mu}$, citation inequality of all-fields case, μ is the mean citation of all articles.
  - For articles with 0 ciation, we follow the convention $0 * log 0 = 0$.
  - $I_1(C)$ is decomposable, $I_1(C) = W + S + IDCP$.

Decomposable Inequality Index

- $W = \sum_\pi \sum_f v^{\pi,f} I_1(c_f^\pi)$, Within-Group term.
  - Weight:$v^{\pi,f}$ is share of total citations received by articles in cell (f, π).
  - For large Π, W is small.

- $S = I_1(\mu^1, ...\mu^\pi, ...\mu^\Pi)$, Between-Group term.
  - Each paper is given the mean citation of articles in its own quantile.
  - Citation inequality is due to articles belonging to different quantiles, i.e. skewness of the all-fields citation distribution.

- $IDCP = \sum_\pi v^\pi I_1(\mu_1^\pi, ...\mu_f^\pi, ...\mu_F^\pi)$.
  - Each paper is given the mean citation of articles in its own cell. Within any quantile π, citation inequality is due to the differences in citation practices across fields.
  - Weight: v^π is the share of total citations received by articles in quantile π.

Decomposable Inequality Index

- $W = \sum_\pi \sum_f v^{\pi,f} I_1(c_f^\pi)$, Within-Group term.
  - Weight:$v^{\pi,f}$ is share of total citations received by articles in cell (f, π).
  - For large Π, W is small.

- $S = I_1(\mu^1, ...\mu^\pi, ...\mu^\Pi)$, Between-Group term.
  - Each paper is given the mean citation of articles in its own quantile.
  - Citation inequality is due to articles belonging to different quantiles, i.e. skewness of the all-fields citation distribution.

- $IDCP = \sum_\pi v^\pi I_1(\mu_1^\pi, ...\mu_f^\pi, ...\mu_F^\pi)$.
  - Each paper is given the mean citation of articles in its own cell. Within any quantile π, citation inequality is due to the differences in citation practices across fields.
  - Weight: v^π is the share of total citations received by articles in quantile π.

Decomposable Inequality Index

- $W = \sum_{\pi} \sum_{f} v^{\pi,f} I_1(c_f^{\pi})$, Within-Group term.
  - Weight:$v^{\pi,f}$ is share of total citations received by articles in cell (f, π).
  - For large Π, W is small.

- $S = I_1(\mu^1, ...\mu^{\pi}, ...\mu^{\Pi})$, Between-Group term.
  - Each paper is given the mean citation of articles in its own quantile.
  - Citation inequality is due to articles belonging to different quantiles, i.e. skewness of the all-fields citation distribution.

- $IDCP = \sum_{\pi} v^{\pi} I_1(\mu_1^{\pi}, ...\mu_f^{\pi}, ...\mu_F^{\pi})$.
  - Each paper is given the mean citation of articles in its own cell. Within any quantile π, citation inequality is due to the differences in citation practices across fields.
  - Weight: v^{π} is the share of total citations received by articles in quantile π.

Data

- Only research papers. One paper is assigned into only one field.
- About 4.4 million articles pubilshed in 1998-2003.
- A common five-year citation window for every year, about 35 million citations.
- 22 broad fields: 20 for natural sciences and 2 for social sciences, distinguished by Thomson Scientific.

Results with Raw Data

Table 1. Citation Inequality Decomposition for Different Inequality Indices and Different Quantile Choices

Choice of	Within-group Term, W (1)	Skew. of Sc. Term, S (2)	*IDCP* Term (3)	Total Citation Inequality. (4)	Percentages In %: (1)/(4) (5)	(2)/(4) (6)	(3)/(4) (7)
10	0.0940	0.6636	0.1179	0.8755	10.7	75.8	13.46
50	0.0300	0.7244	0.1211	0.8755	3.4	87.2	13.83
100	0.0192	0.7348	0.1215	0.8755	2.2	83.9	13.88
1,000	0.0046	0.7488	0.1221	0.8755	0.52	85.53	13.95

Exchange Rates

- Define "exchange rate", $e_f(\pi)$, for articles in cell (f, π):
  - $e_f(\pi) = \frac{\mu_f^\pi}{\mu^\pi}$, how many citations for an article at quantile π of field f are equivalent on average to one citation in all-fields case (the reference situation).
  - If $e_f(\pi)$ varies dramatically across π within a field, no common factor of all quantiles can be estimated.
  - Empirically, $e_f(\pi)$ remain sufficiently constant over quantiles $[706th, 998th]$. 60-70% of citations in each field.

Exchange Rates

- Define "exchange rate", $e_f(\pi)$, for articles in cell (f, π):
  - $e_f(\pi) = \frac{\mu_f^\pi}{\mu^\pi}$, how many citations for an article at quantile π of field f are equivalent on average to one citation in all-fields case (the reference situation).
  - If $e_f(\pi)$ varies dramatically across π within a field, no common factor of all quantiles can be estimated.
  - Empirically, $e_f(\pi)$ remain sufficiently constant over quantiles $[706th, 998th]$. 60-70% of citations in each field.

Exchange Rates

- Define "exchange rate", $e_f(\pi)$, for articles in cell (f, π):
  - $e_f(\pi) = \frac{\mu_f^\pi}{\mu^\pi}$, how many citations for an article at quantile π of field f are equivalent on average to one citation in all-fields case (the reference situation).
  - If $e_f(\pi)$ varies dramatically across π within a field, no common factor of all quantiles can be estimated.
  - Empirically, $e_f(\pi)$ remain sufficiently constant over quantiles $[706th, 998th]$. 60-70% of citations in each field.

Normalization Factor

- Define an average-based exchange rate (ER) over $[706th, 998th]$:

$$e_f = \frac{1}{\pi^M - \pi_m} \sum_{\pi} e_f(\pi)$$

- Normalize raw citations: $c^*_{fi} = \frac{c_{fi}}{e_f}$

Results after Normalization

Table 3. Total Citation Inequality Decomposition Before and After Normalization: *IDCP* Interval Detail

	Within-group Term, *W*	Skew. of Sc. Term, *S*	*IDCP* Term	Total Citation Inequality	Percentages In %:		
	(1)	(2)	(3)	(4)	(1)/(4) (5)	(2)/(4) (6)	(3)/(4) (7)
A. RAW DATA							
All Quantiles	0.0046	0.7488	0.1221	0.8755	0.53	85.52	13.95
[1, 705]			0.0449				5.13
[706, 998]			0.0717				8.18
[999, 1000]			0.0056				0.64
B. EXCHANGE RATE NORMALIZATION							
All Quantiles	0.0051	0.7788	0.0167	0.8006	0.63	97.28	2.09
[1, 705]			0.0127				1.59
[706, 998]			0.0018				0.23
[999, 1000]			0.0022				0.27

Thank you!