Wikipedia in the World of Global Gender Inequality Indices: What The Biography Gender Gap Is Measuring

Max Klein
max@notconfusing.com

ABSTRACT

While Wikipedia's editor gender gap is important but difficult to measure, its biographical gender gap can more readily be measured. We correlate a Wikipedia-derived gender inequality indicator (WIGI), with four widespread gender inequality indices in use today (GDI, GEI, GGGI, and SIGI). Analysing their methodologies and correlations to Wikipedia, we find evidence that Wikipedia's bias in biographical coverage is related to the gender bias in positions of social power.

Categories and Subject Descriptors

J.4 [**SOCIAL AND BEHAVIORAL SCIENCES**]: Sociology; K.6.2 [**MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS**]: Installation Management—*Performance and usage measurement*

Keywords

data mining, Wikidata, Wikipedia, gender gap, demographics

1. INTRODUCTION

Encylopedias have long contained gender bias, both in their *editorship* and in their *biographical coverage*. [8] [7]. Like its historical counterparts, Wikipedia has been estimated to have a female authorship of around only 13% - 16%[3]. A skew of the same order of magnitude has been found by studies of biographical articles in Wikipedia [5], [1], [7]. More recently Wagner et. al has compared gender statistics derived from Wikipedia and gender inequality indices, concluding "to a certain extent gender inequalities of the real world manifest on Wikipedia"[9]. Yet it is important to disentangle biases that exist in the larger world and those that are introduced by Wikipedians and Wikipedia policy. We seek to understand in more detail which social factors that gender inequality indices measure are also reflected in Wikipedia.

OpenSym '15 San Francisco, Calif. USA

2. RESEARCH QUESTION

RQ1: Using other gender inequality indices as a basis, which social factors does Wikipedia biography inclusion measure?

3. METHOD

We correlate four general gender inequality indices to a Wikipedia-derived indicator, and then compare which factors in the highest and lowest correlated indices are similar.

Our Wikipedia-derived indicator is called the *Wikipedia Gender Inequality Indicator* (**WIGI**), and utilizes Wikidata. Wikidata is a database that stores and feeds semantic facts to every language edition of Wikipedia. Its October 14 2014 dataset included a total of 1,061,634 Wikidata items about humans with a *date of birth* and a *citizenship* or a *place of birth* property that is a country. Place of birth and citizenship we recognize may not be the same, but in case of contradiction we simply select one at random. As we are looking at large aggregates we don't believe this will have a significant effect. Ultimately the *ratio of female and nonbinary gendered biographies* to total biographies the score for a country[4]. WIGI that list of countries along with these computed biography gender ratios.

Of the other gender inequality indices there is a consensus that the complexity of the issue means that no single index is the best [6] [2]. We use four "general" indices that are in widespread use.

- The Gender Development Index (**GDI**) from the United Nations Development program "measures gender gap in human development achievements in three basic dimensions of human development: health, measured by female and male life expectancy at birth; education, measured by female and male expected years of schooling for children and female and male mean years of schooling for adults ages 25 and older; and command over economic resources, measured by female and male estimated earned income." [1]

- In Education, Social Watch's Gender Equity Index (**GEI**) "looks at the gender gap in enrolment at all levels and in literacy; economic participation computes the gaps in income and employment and empowerment measures the gaps in highly qualified jobs, parliament and senior executive positions." [2]

[1] `http://hdr.undp.org/en/content/gender-development-index-gdi`
[2] `http://www.socialwatch.org/node/14366`

- The Global Gender Gap Index (**GGGI**) was developed by the World Economic Forum in 2006. The GGGI is intended to allow comparison of gender gap across different countries and years, it focuses on four areas: economic participation and opportunity, educational attainment, political empowerment and health statistics.

- The Social Institutions and Gender Index (**SIGI**) is of the OECD Development Centre from 2007. A composite indicator of gender equality that solely focuses on social institutions (norms, values and attitudes), SIGI uses the five dimensions of discriminatory family code, restricted physical integrity, son bias, restricted resources and assets, and restricted civil liberties.

In each index higher scores represent higher equality, however in order to normalise comparisons we focus only on the positional rank of the countries, not the precise score. To understand how close two indices are we use the Spearman rank correlation coefficient.

We next compute a calibration step on the year of birth of Wikipedia biographies to maximise the correlation between WIGI and a general index. The end date for biography inclusion is held constant at the present date. That is, we are looking at what time until present the biography data of Wikipedia makes a by-country index that most matches a general index.

Finally for each general index we compute the maximum Spearman rank correlation along with the start decade that achieves it. Full data[3] and code[4] available online.

4. RESULTS

We produce a comparison table of indices, their correlation, the correlation significance, and the maximizing start decade, ordered by correlation, in table 1. Each general index shows some statistically significant moderate correlation with WIGI. In the end the GEI most highly correlates with WIGI with Spearman rank correlation 0.417, and the GDI the least at 0.278.

Also we find that each general index most highly correlates with WIGI around 1910. Intuitively this makes sense in light of the fact that general indices are more a measure of modern history, as they seek to measure the present day. The present day humans are those born about 1910 or later. Wikipedia has information dating back past 1000 BCE, but this information is not useful in describing the world that the general indices do.

5. DISCUSSION

How do the methodologies of the GEI and GDI show what WIGI is measuring? Looking back at their descriptions we find that both indices share overlapping features in *education* and *income*, but diverge on the features of *empowerment* and *health*. In their similarities of education and income we find related measurements of school enrolment, years of schooling and earned income. The disparity that emerges is that GEI additionally measures empowerment by positions of power whereas GDI additionally measures life expectancy. Since GEI is the is the most similar to WIGI and GDI is the least similar, this suggests that the WIGI is more highly correlated to women's positions of power by country than to life expectancy by country.

Table 1: WIGI's correlation to general indices and calibrated start date.

Index	Spearman Correlation	Significance	Calibrated Start Decade
GEI	0.417	$p<0.001$	1910
SIGI	0.338	$p<0.001$	1910
GGGI	0.310	$p=0.03$	1890
GDI	0.278	$p<0.001$	1910

That positions of power bias is commensurate with what we would expect from Wikipedia's notability policies. Notability in Wikipedia, although varying by language, essentially defers to inclusion in the journalistic record. That means that humans in positions of power, as GEI imports, would would have articles in Wikipedias in greater proportion, because more powerful positions are more covered in media. Yet, because of the imperfect correlation we find with the GEI, we cannot say that Wikipedia's gender bias is entirely due the social positions of power bias - it is only one factor.

Finally we conclude that Wikipedia's biographical bias is closer to the gender bias in highly-qualified jobs, political and executive positions than longevity.

6. REFERENCES

[1] Y.-H. Eom, P. Aragon, D. Laniado, A. Kaltenbrunner, S. Vigna, and D. L. Shepelyansky. Interactions of cultures and top people of wikipedia from ranking of 24 language editions. *PLoS One*, 2014.

[2] A. Hawken and G. L. Munck. Cross-national indices with gender-differentiated data: What do they measure? how valid are they? *Social Indicators Research*, 2011.

[3] A. S. Hill, Benjamin Mako. The wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PLoS ONE*, 2013.

[4] M. Klein and P. Konieczny. Gender gap through time and space: A journey through wikipedia biographies and the "wigi" index. http://arxiv.org/abs/1502.03086.

[5] S. Lam, A. Uduwage, Z. Dong, S. Sen, D. Musicant, L. Terveen, and J. Riedl. Wp:clubhouse? an exploration of wikipedia's gender imbalance. In *Wikisym'11*.

[6] M. Mills. Gender roles, gender (in)equality and fertility: An empirical test of five gender equity indices. *Canadian Studies in Population*, 2010.

[7] J. M. Reagle and L. Rhue. Gender bias in wikipedia and britannica. *International Journal of Communication*, 2011.

[8] G. Thomas. *A Position to command respect: Women and the Eleventh Britannica.* The Scarecrow Press, Meutchen, NJ, 1992.

[9] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *CoRR*, abs/1501.06307, 2015.

[3] `https://github.com/notconfusing/WIGI/blob/master/helpers/foreign_indexes/WIGI_comparison.csv`

[4] `http://nbviewer.ipython.org/github/notconfusing/WIGI/blob/master/World%20Economic%20Forum%20Comparison.ipynb`