Methodological supplement to "Reducing Bias in Wikipedia's Coverage of Political Scientists"

My methodology for counting bias on Wikipedia is straightforward: I rely on Wikipedia's own system of categories, which are tags that classify every page according to its main attributes, and I simply tally the number of articles in each category. I also validated that the main categories were applied to every appropriate page, adding them where necessary. I offer three caveats about this approach.

First, I consider any attempt to define "political scientist" to be beyond the scope of my attempts to measure bias. I simply take Wikipedia's category of political scientists to contain exactly the set of Wikipedia pages that are biographies of political scientists.

Second, in simple counting exercises, I rely on Wikipedia categories and lists. In the case of political scientists, unfortunately the only traits that are tracked are nationality and gender. To tally the proportions of biographies about political scientists that belong to those categories, first I collected the titles of English Wikipedia pages which belong to the category "political scientist". Because pages can be placed in a category multiple times, I next reduced that to a list of only unique article titles. This produced a list of 3,335 unique names (at the time of data collection in late 2020). I then manually checked each of these 3,335 articles to ensure that every page about a woman was included in the category "women political scientists". Controversially, it has long been standard practice in some areas of Wikipedia to have a special sub-category for women but not for men (Filipacchi 2013). So, while I ensure that all the pages about women are included in the "women political scientists" category, there is no analogous "men political scientists" category to sort most of the remaining pages into.

The counts of political scientists by nationality are simply counts of the number of political scientists in each category by nationality. So for example, the number of American political scientists is obtained by taking the number of political scientists in the category "American political scientists" and all of its subcategories (just those subcategories which are marked as "non-diffusing", to avoid double-counting). Note that this figure is lower than the total number of political scientists multiplied by 0.44. The reason is that not all biographies of political scientists are categorized according to nationality. Americans make up 44% of the political scientists with biographies that are categorized by nationality, and I assume that the distribution of nationalities among the uncategorized pages is the same as the distribution of nationalities among categorized pages. Of course that could be wrong; if we believe that pages about Americans have higher article quality overall, then we might imagine that the probability of a page about an American remaining uncategorized by nationality is lower than that probability for a page about a non-American, which would mean that the 44% figure overstates the bias towards Americans. Or we might think the opposite, perhaps because Wikipedia is so US-centric that editors feel more of a need to explicitly identify non-Americans while treating Americans as the default. Such systematic differences are possible, but the effect cannot be large; because we know how many pages are uncategorized, we can compute an error range. In the extreme case that every uncategorized biography is about a non-American, the proportion of Americans overall would be about 35%. In the opposite extreme, if every page without a nationality category were about an American, the proportion of Americans overall would be about 54%. So the true proportion of Americans among the subjects of Wikipedia's political scientist biographies is certainly within the range [0.35; 0.54]. If we assume that pages without nationality categories have a similar proportion to categorized pages, it is close to 0.44.

The third caveat is that I make two crude assumptions in order to count gender proportions. First, I assume that I can identify the gender of a political scientist by the pronouns which are used to refer to them on their Wikipedia page; Wikipedia has a policy of honouring each page subject's preferred pronouns, so for a sufficiently well-maintained page this assumption should hold. In cases where the subject does not have a Wikipedia page (this is only relevant when I am counting proportions among cited authors or authors of recommended works), I search for material that was written by them or their employer, and consult the pronouns used there. The pronoun method remains imperfect because not all people referred to by the pronouns "she" and "her" are women, and not all people referred to by the pronouns "he" and "him" are men, but I assume this is a negligible source of error. Second, I am not precisely classifying "man" or "woman", I am classifying "woman" or "not woman" (as measured by use of she/her pronouns). Importantly, there are some biography subjects or cited authors who use they/them pronouns; these pages are not part of the category "women political scientists" and they are not included in my count of the number of women. Changing this rule would not affect the percentages stated.

REFERENCE

Filipacchi, Amanda. 2013. Wikipedia's Sexism Toward Female Novelists. *The New York Times*. https://www.nytimes.com/2013/04/28/opinion/sunday/wikipedias-sexism-toward-female-novelists.html