Quantifying the Gap: A Case Study of Wikidata Gender Disparities

Charles Chuankai Zhang
zhan6914@umn.edu
GroupLens Research
University of Minnesota
Minneapolis, Minnesota, USA

Loren Terveen
terveen@umn.edu
GroupLens Research
University of Minnesota
Minneapolis, Minnesota, USA

ABSTRACT

Much prior research has found *gender bias* in peer production systems like Wikipedia and OpenStreetMap. This bias affects both women's participation in these platforms and content about women on these platforms. We investigated the gender content gap in Wikidata, where less than 22% of items that represent people are about women. We asked: what is the *source* of this bias? Specifically, does it *originate from the actions of Wikidata editors* or *from external factors*; that is, does it simply reflect existing real world gender bias? We conducted a quantitative case study that found: (i) the most popular categories of people included in Wikidata represent male-dominant professions, such as American football; (ii) within a selected set of professions where we could obtain gender distribution data, Wikidata is no more biased than the real world: men and women are included at similar percentages, and the quality of items representing men and women also is similar. We provide possible explanations for our findings and implications for addressing the Wikidata content gap.

CCS CONCEPTS

• **Human-centered computing → Wikis**.

KEYWORDS

Wikidata, peer-production, structured data

ACM Reference Format:
Charles Chuankai Zhang and Loren Terveen. 2021. Quantifying the Gap: A Case Study of Wikidata Gender Disparities. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3479986.3479992

1 INTRODUCTION

Wikidata is a Wikimedia project that serves as "a free and open knowledge base that can be read and edited by both humans and machines".[1] It stores structured data about real world objects and concepts. It is widely used both by Wikimedia projects – particularly Wikipedia – and many other sites and services, such as Google, Quora and Musicbrainz. One exemplary use of Wikidata in Wikipedia is to provide data for *infoboxes*; Figure 1 shows an example infobox and its corresponding Wiki markup, which indicates data to be fetched from Wikidata.

Given Wikidata's role as a knowledge repository used by a variety of different sources, questions about the nature and quality of its information are important. Any problems or biases in Wikidata's representation of knowledge may be propagated to search engines, question answering platforms, or other online communities that use Wikidata as a source of reference or ground truth. We are particularly interested in an issue that plagues many peer production communities including (as detailed below) Wikipedia and OpenStreetMap: *gender*[2] *bias*, the under representation of content about women compared to content about men. In line with previous studies, we found that only 22% of Wikidata items that represent people are about women. The goal of our research is to begin to identify the *source of this bias*. Identifying the source of bias is important because different sources may require different remedies.

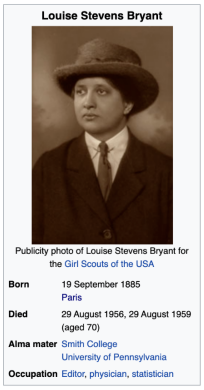
Most generally, bias can originate *from the actions of Wikidata editors* or *from external factors* (or obviously, from a combination of the two). Perhaps Wikidata editors tend to add items about men proportionally more often than items about women; or more subtly, maybe they favor adding content about categories of people where men dominate (say, popular American sports) rather than where women dominate or the genders are equally represented. On the other hand, perhaps the manifest gender bias in Wikidata content merely reflects real world biases. That is, due to discrimination and systemic biases, maybe women are underrepresented in the kinds of activities and achievements that lead people to be considered "notable" enough to be represented in Wikidata: "The entity must be notable, in the sense that it can be described using serious and publicly available references[3]." This article defines several other criteria to determine whether an entity is acceptable for representation in Wikidata. However, at least for items representing people, our reading is that notability clearly is required. Therefore, since inclusion in Wikidata is presumed to indicate that a person is "notable" (and lack of inclusion may indicate that a person is **not** "notable"), gender biases in Wikidata can lead consumers of Wikidata to form incorrect perceptions about the comparative notability of women and men.

[1] https://www.wikidata.org/wiki/Wikidata:Main_Page

OpenSym 2021, September 15–17, 2021, Online, Spain

ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479992

[2] As we detail below, nearly all Wikidata items that represent people have a gender of either female or male, so in this paper we consider only these two genders.
[3] https://www.wikidata.org/wiki/Wikidata:Notability




```
{{Infobox person/Wikidata |
fetchwikidata=ALL|suppressfields=citizenship}}
```

Figure 1: An example of Wikipedia infobox powered by Wikidata

As we prepared to investigate these possibilities, we realized that we needed to think about the representation of men and women *within specific professions* as people are likely to be recognized as *notable* for the accomplishment in the field they work on. In particular, we needed two types of data:

- The overall distribution of men and women in different professions; for this, we used United States[4] Bureau of Labor Statistics.
- The assessed "notability" of individuals in different professions; for this, we used lists of award winners for a selected set of professions.

After obtaining these datasets, we could formulate a guiding **research question** for this study:

> *To what extent is Wikidata **reflecting real world gender bias** vs. **introducing additional gender bias**?*

In brief, we found that:

- Wikidata editors "over sample" male-dominated professions such as American football and baseball.
- However, within a selected set of professions for which we obtained overall gender distributions and external "notability" assessments, Wikidata gender distribution is no more biased than the real world.
- Moreover, the quality of Wikidata items representing women and items representing men are equivalent.

The rest of this paper is organized as follows. First, we introduce and illustrate some key Wikidata concepts. Second, we summarize related work focusing on gender disparities in online communities. Third, we elaborate on our analytic framework and describe our data and methods. We then present our results, and conclude with a discussion of the implications of our results for future research and remedies to the observed gender gap.

[4]Analyzing only United States data is a limitation of our study. We explain why we did this in the section *Data and Methods*

2 WIKIDATA CONCEPTS AND TERMINOLOGY

Wikidata objects that represent real world entities like people are called *items*. *Label* is the most common name that an item is known by, and *description* is a short piece of information that describes the item[5]. For every item, the major bulk of its information and characteristics are stored as a list of *Statements*. A statement consist of two parts: a *claim* that the item has certain characteristics and a list of references that back up this claim. The most common form of a claim is a *property-value* pair that assigns one or multiple values to a certain property. For example, the Wikidata object[6] referred to in Figure 1 includes the following statements:

instance of **human**
sex or gender **female**
occupation **publicist**
writer
editor
physician
statistician
medical writer

The example above has three claims. The first two claims are formed by one-property-one-value pairs while the third claim is an occupation property that pairs with six values.

In this paper, we analyze only Wikidata items that are *instance of* **human**, i.e., items directly representing people. There are other types of content that could be considered to be "related" more directly to either men or women (for example, see [26] and [31]). However, since we are using real world data about gender composition of people within professions as a baseline for comparison, for our purposes it makes sense to consider only Wikidata items that represent people.

3 RELATED WORK

We report here on our research using quantitative methods to investigate the sources of gender bias in the Wikidata peer production system. Thus, our discussion of prior research focuses on work in this context that applies similar methods. However, it is worth noting that other work takes a more conceptual and theoretical approach to the issue of gender bias in online communities, including peer production systems. Some of the most relevant strands of this work [10, 32] draw on feminist HCI [3] to critique the underlying epistemological and procedural foundations of communities like Wikipedia. This work is not directly relevant to our current study, but it offers alternative perspectives aimed to create a more pluralistic and inclusive community and content, thus addressing gender bias at a fundamental level.

Prior research on peer production systems has found significant gender gaps in participation and in content coverage. Much work has focused on Wikipedia. By 2010, studies had begun to appear that found that women were a small minority of Wikipedia contributors. Glott et al. [13] conducted a survey finding that less than 13% of Wikipedia editors are women (although a revised analysis suggested that the number might be around 16% [20]). Several

[5]Since Wikidata is language independent, an item can have labels and descriptions in multiple languages.
[6]https://www.wikidata.org/wiki/Q24455644

quantitative studies found similar results and added a further finding: women were even underrepresented among the most active Wikipedia editors [1, 26]. Cabrera et al. [5] found a gender gap in participation in article talk pages, where issues concerning article content are raised and discussed. Hargittai and Shaw [18] summarized panel survey data to conclude that the most likely contributors to Wikipedia are highly skilled men. Gender gaps in participation have been attested in other populations like OpenStreetMap contributors [9, 11], StackOverflow contributors [36, 39] and open source software developers [12].

What factors lead to this large participation disparity? Prior research discovered both internal and external reasons. First, there is a significant difference between men and women in terms of online behavior. Iosub et al. [22] suggests that women Wikipedia contributors communicate in a more social and emotional manner and that women contributors are more relationship-oriented. Laniado et al. [27] found that women editors tend to communicate in a more positive tone. It is supported by the finding that distaste of high level of debate in contribution process and certain tasks like deletion are also reasons why women editors turn away from Wikipedia [4, 7]. Meanwhile, external factors are also examined in order to understand how the environment and culture of a platform contributes to the gender disparity in editors. Through an interview study, Menking and Erickson [30] found that women editors avoid certain kinds of areas or tasks that involve too much drama and stress. Organizational tensions in sociocultural norms may also cause Wikipedia women editors to experience isolation and emotional exhaustion [8]. Lir [29] analyzed the participation process and discovered pre-edit and post-edit barriers that deters women from contributing to Wikipedia.

Generally, gender gaps are a consequence of the *culture, dynamics, and values of online communities* [33]. The various types of gender gaps cause different types of harms. A *contributor gap* often leads to a *content gap* since women and men overall may differ in their interests and specializations [6, 9], and thus the types of content they create and edit. Specifically, previous research showed that Wikipedia's editor gender gap was associated with poor coverage and quality of topics that appealed more to women than men [26]. Other research found that Wikipedia biographies covered a much higher proportion of men than women, but the women who were included tended to be more notable than men, due to a hypothesized "glass ceiling" effect [38]. This research also found that articles about men and women covered different types of topics; for example information about relationship and family was more likely to be included in Wikipedia articles about women [37], while cognition related content was more likely to be included in articles about men [14]. In addition, OpenStreetMap and Google MapMaker both were shown to have gender biases in the types of places they included [34].

In previous research on gender disparity, Wikidata was used as a data source for measuring disparity in Wikipedia. For example, Klein et al. [24] built a Wikipedia gender gap indicator using Wikidata as a data source. An in-depth analysis of claim coverage and Wikidata human items by place of birth and citizenship was conducted to help them build up the indicator. In a case study on members of the European Parliament, Hollink et al. [21] compared the number of claims and family/relationship related properties between men and women Wikidata items. They found only a small difference in number of properties. They also found no evidence indicating family/relationship related properties shows up more in Wikidata items representing women, in contrast to the result from Wikipedia.

Thus far there is no systematic account of the gender gap in Wikidata; specifically, there has not been an investigation of the causes of the gap. This is important because different causes may require different solutions. If the gap originates from actions of Wikidata editors, then solutions would have to focus on the editors, for example, the composition of the editor population or tools designed to change editor behavior. On the other hand, if the gap primarily reflects existing real world biases, then solutions might require new policies to "over sample" external data to fight against these biases.

4 DATA AND METHODS

As we have explained, the Wikidata gender gap could originate *from the actions of Wikidata editors* or *from external factors*. In other words, are Wikidata contributors *causing the gap* or *reflecting an existing gap*? To answer this question, we need to compare Wikidata data to external data.

We realized we needed data organized by *professions*: as Kay et al. [23] noted: the "portrayal of occupations" is a "topic of societal importance that has recently received attention and efforts to ameliorate biases". Moreover, different professions have different gender distributions and different barriers to advancement, that is, what types of people become recognized as "notable". Therefore, organizing data by profession let us address several specific questions:

(1) *How does the Wikidata gender distribution within a profession compare to the overall gender distribution within that profession?* To answer this question, we need a dataset of gender distribution by profession.
(2) *How does the Wikidata gender distribution within a profession compare to that profession's "notability" assessments?* To answer this question, we need a dataset of people recognized as notable within various professions, along with the gender distribution of the people so recognized.
(3) *Which types of professions have most coverage in Wikidata? Are these professions more balanced or biased in gender representation?* To answer this question, we need a dataset of Wikidata items that represent people, where the person's profession and gender also are provided.

We faced several challenges in collecting the datasets that led us to take an iterative approach to defining and then *refining* the datasets. We narrate these challenges and explain the assumptions we made as we describe each dataset.

4.1 Gender Distribution by Profession: BLS Dataset

For overall gender distribution within professions, we used the United States Bureau of Labor Statistics' Current Population Survey

dataset[7] as of the year 2019. For our purposes, this let us calculate the gender distribution[8] within a large number of professions.

Analyzing only United States data is a limitation. We accepted this limitation due to the availability of a high-quality data source, which is not duplicated globally. Further, this dataset has been used in previous research [23, 24] to serve as a ground truth of gender representation, again with analysis limited to the United States.

4.2 Gender Distribution by Profession: Wikidata Dataset

4.2.1 Initial Dataset Construction. We used the October 19, 2019 Wikidata data dump. We first extracted all Wikidata items that represent people, that is, items whose *instance of* (P31) property had the value *human* (Q5). This resulted in 5,477,414 items, comprising 8.5% of all items in the dump. We next filtered to include only items that had values for four properties necessary for our analysis: *sex or gender* (P21), *occupation* (P106), *date of birth* (P569) and *country of citizenship* (P27). This left us with 2,513.518 items, or just under 46% of all the human Wikidata items. We further required certain values for these properties:

- *country of citizenship* must be *United States of America* (Q30); this was necessary for comparison with the BLS dataset.
- *date of birth* (P569) had to be at least as recent as 1950; we did this for comparison with the two external datasets, as most people of this age are still employed (and thus represented in the BLS data) and have had the opportunity to become recognized as "notable" in their profession.[9]

This final filter let us with a dataset consisting of 141,562 Wikidata items representing people with US citizenship, born after 1950, with a known gender and profession.

4.2.2 Organizing by Profession. We next had to group the items in this dataset by *profession*. We initially limited ourselves to professions with more than 100 items; this yielded 133 professions. These professions and their counts are listed in the Appendix. Next, we needed to match those professions to the profession listed in the census dataset. Like others before us [23], we encountered the problem of polysemy; many BLS categories cover multiple distinct professions that are distinguished in Wikidata. For example, the BLS 'Athletes, coaches, umpires, and related workers' profession corresponds to 42 distinct Wikidata professions, such as "American football player", "baseball player", "basketball official", and "sports commentator". Only seven of the 133 Wikidata professions with at least 100 items had a 1-to-1 match with BLS categories. Five of these seven professions were academic professions: chemistry, computer science, economics, psychology, and sociology. We selected these academic professions for our notability dataset to create a focused baseline for comparison.

However, we still had one more step for the five selected academic professions. Some Wikidata professions may be subclasses of others, represented using the *subclasses of* (P279) property. For example, a *theoretical chemist* (Q85519878) is a subclass of *physical chemist* (Q16744668) and *physical chemist* is a subclass of *chemist*. Thus, someone whose profession is *theoretical chemist* should be included among *chemists* in our dataset. Therefore, for each of the five selected professions we expanded the subclass hierarchy until we reached leaf nodes. Then for each human item in our dataset, if its *occupation* property included any profession in the class hierarchy rooted at one of the selected professions, we assigned the item to that profession.

Table 1: Academic professions and their corresponding associations

Profession	Society & Association
chemist	American Chemical Society[11]
psychologist	American Psychological Association[12]
sociologist	American Sociological Association[13]
computer scientist	Association for Computing Machinery[14]
economist	American Economic Association[15]

4.3 Notability Dataset for Five Academic Professions

Finally, since people are supposed to be notable to be represented in Wikidata, we needed to obtain external datasets of notable people within the five selected academic professions. We believe that professional society's award recipients are the best source for this. To be clear, we are not saying anything about whether this type of recognition is fair or unbiased; we simply are saying that it reflects a profession's assessment of the notable people within its field.

Table 1 lists for each of the five selected academic professions the professional society from which we obtained lists of award recipients. We collected this information in September 2019. We wanted to "synchronize" the notability datasets with Wikidata and BLS datasets. Recall that the BLS dataset deals with currently employed people, and we limited the Wikidata dataset to people born after 1950. We decided that by the time people were 30, they were almost certain to be employed, and they had some chance of having received recognition in their field. Therefore, we included only award recipients from the professional societies who received their award beginning in 1980.

Finally, we determined the gender of award recipients in two ways. First, if an award recipient was included in Wikidata, we retrieved their gender from Wikidata. (We report Wikidata coverage of the notability datasets below.) Otherwise, we used the gender-guesser python library[10] which has a 97.34% gender identification accuracy on Wikidata dataset [25]. The result of this tool for any given name will be one of *unknown (name not found), andy (androgynous), male, female, mostly_male, or mostly_female*. We used this tool and kept only the *male* and *female* classification results. The rest of the data were hand labeled using different sources such as Google and Wikipedia. We realize that this procedure may make incorrect gender classifications, and this is a limitation of our approach.

[7] https://www.census.gov/programs-surveys/cps.html

[8] as noted previously, we limited ourselves to men and women genders due to data availability issues.

[9] We also observed that some items do not have an exact birth date. For example, some people are listed only as born in the "*20th century*"; in this case, the data in the dump is *+2000-00-00T00:00:00Z*. We filter out these items, too.

[10] https://pypi.org/project/gender-guesser/

Table 2: List of genders and their proportion in Wikidata dataset and 5 profession dataset

Wikidata QID	Gender	Wikidata	5 Professions
Q6581072	female	24.2%	19.4%
Q6581097	male	75.6%	80.5%
Q1052281	transgender female	0.10%	0.06%
Q2449503	transgender male	0.003%	0
Q48270	non-binary	0.002%	0
Q301702	two-spirit	7e-6	0
Q1097630	intersex	3e-5	0
Q505371	agender	1e-5	0
Q189125	transgender person	7e-6	0

Table 2 shows genders listed in Wikidata and the five selected academic professions and their proportions. *Male* and *Female* genders account for more than 99.8% of the data. Therefore, we only were able to analyze distribution of these two genders. Future work is necessary to obtain sufficient data to examine biases across a wider range of genders.

5 RESULTS AND ANALYSIS

We next present our results, organized around potential Wikidata gaps (relative to external data) in *coverage* and *quality*.

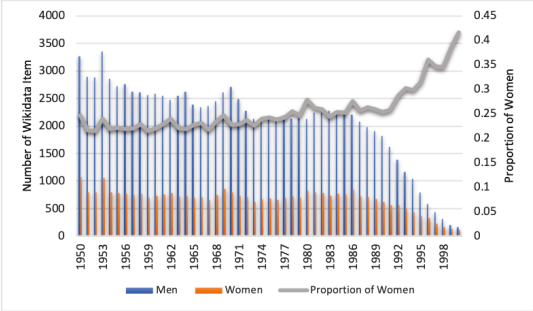


Figure 2: Number of items and proportion of women per year

5.1 Coverage Gap

Figure 2 shows the number of human items and gender proportion per year in the Wikidata dataset. Blue bars represent the number of Wikidata items about men born in each year, and the orange bars represent the number of items about women born in that year. The line on the chart shows the gender proportion trend over time. From the graph, we observed that the proportion of Wikidata items about women ranges between 0.2 and 0.25 for birth years 1950 to 1990 and has increased steadily since then, reaching 0.4178 for the 2000 birth year.

[11]https://www.acs.org/content/acs/en/awards.html
[12]https://www.apa.org/about/awards
[13]https://www.asanet.org/about/awards
[14]https://awards.acm.org/
[15]https://www.aeaweb.org/about-aea/honors-awards

To investigate further, we list the five professions with the most items by five-year blocks. Table 3 shows the top professions every five years and their percentage in the dataset. We can observe that only eight professions appear in all the top five ranking lists. Four of them are sports related professions dominated by men: American football player (99.83%), basketball player (86.98%), baseball player (99.78%) and association football player (85.54%). As for the other professions, politician (76.85%) is also heavily biased towards men, while actor (51.12%), singer (54.41%) and writer (59.13%) are more equally represented in Wikidata. While 1392 professions occurred in the Wikidata dataset, the top five professions cover at least 30% of the data.

Thus, we can articulate an obvious gender coverage bias immediately: many of the professions most commonly represented in Wikidata are male-dominated. This in turn will skew the overall gender distribution in favor of men.

We can make several conjectures concerning the increase in representation of women among people born after 1990. First, the number of people born in this time span included in Wikidata decreases significantly. For example, several thousand people are represented in Wikidata for each 1980s birth year, but fewer than 300 for birth year 2000. This makes sense, as people who are only in their 20s have had less chance to become "notable". Second, among people born in the 1990s who are represented in Wikidata, non-sports related professions – which are much less male dominated – make up a significantly larger proportion. For example, for people born between 1986 and 1990, four of the five top professions are sports related, male-dominated, and they collectively account for nearly 46% of Wikidata human items. The one non-sports profession, *Actors*, which has virtually equal gender distributions, accounted for just under 9% of human items. However, for people born between 1996 and 2000, there are three (male-dominated) sports professions in the top 5, which collectively account for just under 29% of human items, while *Actors* is joined in the top 5 by *Singers* – another profession with close to equal gender distribution – and these two professions together accounted for over 23% of human items.

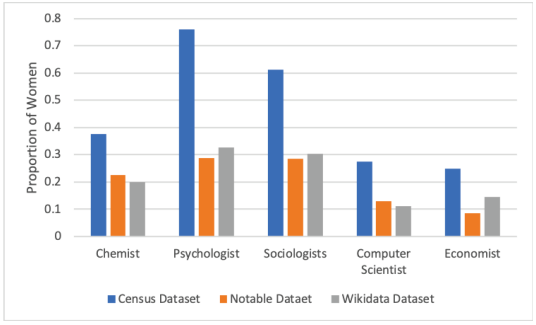


Figure 3: Proportion of women in 5 academic professions across three datasets

We next compared Wikidata gender distribution within the five selected academic professions to the gender distribution in the profession as a whole (BLS data) and in professional societies' notability assessments. Figure 3 shows the proportion of women in the five academic professions in each of our three datasets. We first

Table 3: Top 5 professions with most data in every five years

Year	Top 1	Top 2	Top 3	Top 4	Top 5	Proportion
1951 - 1955	politician	actor	American football player	writer	baseball player	30.5%
1956 - 1960	politician	American football player	actor	baseball player	writer	32.0%
1961 - 1965	American football player	actor	politician	baseball player	basketball player	34.6%
1966 - 1970	actor	American football player	politician	baseball player	basketball player	32.9%
1971 - 1975	American football player	actor	baseball player	politician	basketball player	35.9%
1976 - 1980	American football player	actor	basketball player	baseball player	singer	40.7%
1981 - 1985	American football player	actor	basketball player	baseball player	association football player	47.3%
1986 - 1990	American football player	basketball player	actor	association football player	baseball player	54.8%
1991 - 1995	American football player	basketball player	association football player	actor	baseball player	57.1%
1996 - 2000	actor	association football player	basketball player	American football player	singer	52.2%

observed that both Wikidata and professional societies' notability assessments include a much higher proportion of men than in the profession as a whole. On the other hand, the graph suggests that Wikidata gender distributions are no more biased than the professional societies. There is a high within-pair correlation (r = 0.923) between the gender proportion of notable dataset and Wikidata dataset so we have sufficient statistical power to run a paired t-test on this small sample size (n=5). The t-test result between the Wikidata dataset and the notable dataset shows that there is not a significant difference between them (p = 0.414).

The previous analysis let us compare the overall gender distributions of the Wikidata and the notability datasets. We also examined the *specific* coverage of the notability datasets – that is, the people recognized by the five professional societies – in Wikidata. For every person in the notability datasets, we checked whether they were represented in Wikidata. We used their name as a query, and considered ourselves to have found a match if and only if exactly one human item with a matching profession is found. For example, the 1980 Association of Computing Machinery (ACM) Turing Award winner was Tony Hoare. The ACM lists his name as C. ANTONY ("TONY") R. HOARE[16]. Using this as a query to Wikidata returns exactly one item – *Tony Hoare (Q92602)* – and this item is an *instance of* a *human*, and its *occupation* property includes the value *Computer Scientist*. Therefore, this is a successful match. The 1987 Turing Award Winner was John Cocke[17]. The ACM lists his name as "John Cocke". Using this as a query to Wikidata returns a large number of Wikidata items, but only one – *John Cocke (Q92632)* – is an *instance of* a *human*, and has an *occupation* property that includes the value *Computer Scientist*. Therefore, this too is a successful match. On the other hand, many recipients of some other ACM awards, such as Distinguished Member, are not found in Wikidata at all. A final note is that we once again use the Wikidata profession hierarchy in this process. So for example, if someone honored by the American Chemical Society was listed as a *theoretical chemist* in Wikidata, we treat this as a match, too.

Figure 4 shows the results of this analysis. It shows that there is a large difference in coverage among professions. More than 90% of economists recognized by the American Economic Association are represented in Wikidata, while only about 40% of chemists recognized by the American Chemical Society are included. In no profession is there a significant difference between the proportion of men and women award recipients who are represented in Wikidata. In absolute terms, in four of the five professions, a higher proportion of women award recipients is included in Wikidata.

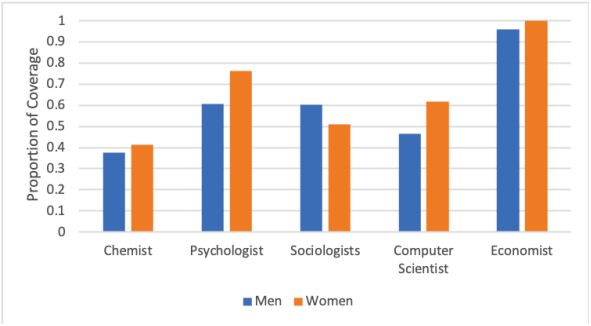


Figure 4: Notable dataset Coverage in Wikidata

5.2 Quality Gap

In this section, we investigate whether there is a difference in the *quality* of Wikidata items that represent men and items that represent women. We limit our analysis to human items within the five academic professions since items within these professions are likely to be characterized by a similar set of properties. In other words, it is easier to compare the quality of two academic professionals than to compare an academic professional to an actor or athlete. In our analysis, we use both specific quality metrics and other factors that are associated with quality.

5.2.1 Direct Metrics. We use the Objective Revision Evaluation Service (ORES) [17] to evaluate the quality of Wikidata items. ORES is a service provided by the Wikimedia Foundation that predicts edit quality and assists content moderation for various Wikimedia projects [16].

A Wikidata data dump includes the most recent revision of each Wikidata item at the time the dump was created. We extracted revision IDs from the Wikidata dataset and used ORES' quality evaluation API to estimate item quality. ORES uses the Wikidata quality assessments, which range from A (highest) to E (lowest)[18]. Specifically, ORES returns the probability that an item should be classified at each level. We then used the weighted sum formula (Formula 1) proposed by Halfaker [15] to compute a single score ranging from highest quality (4) to lowest (0). An item would be

[16]https://amturing.acm.org/award_winners/hoare_4622167.cfm
[17]https://amturing.acm.org/award_winners/cocke_2083115.cfm
[18]https://www.wikidata.org/wiki/Wikidata:Item_quality

scored a 4 if ORES predicted a 100% probability that the item should be classified as quality level A, and would be scored a 0 if ORES predicted a 100% probability that the item should be classified as quality level E. Figure 5 shows an an example of a prediction given by ORES to a particular item; using the weighted sum formula, the item would receive a score of 2.9901, which corresponds to quality level B.

$$\begin{aligned}\text{Weighted Sum} = {} & 4 \times \text{P(item is of quality A)}+ \\ & 3 \times \text{P(item is of quality B)}+ \\ & 2 \times \text{P(item is of quality C)}+ \\ & 1 \times \text{P(item is of quality D)}+ \\ & 0 \times \text{P(item is of quality E)}\end{aligned} \tag{1}$$


```
"score": {
  "prediction": "B",
  "probability": {
    "A": 0.07928625288897875,
    "B": 0.8437658868814822,
    "C": 0.06826526593440399,
    "D": 0.005158008654169078,
    "E": 0.0035245856409659073
  }
}
```


Figure 5: An example of ORES quality score prediction

As shown in the leftmost part of Figure 6 (and confirmed by the t-test result in Table 4), there is no significant difference in the ORES scores between Wikidata items representing women and items representing men. However, we want to take a closer look at a few specific important features, to see if any of them exhibited significant differences between men and women items.

- The number of *claims* constitutes the total amount of information about a Wikidata item.
- *Labels* and *descriptions* are multilingual, so the more of each, the better the representation of an item in multiple languages.
- *Sitelinks* link to other Wikimedia projects, so more sitelinks means the item is better connected to the larger Wikimedia ecosystem.

The remainder of Figure 6 shows box plots four ORES scores. We performed an independent t-test on these features. Table 4 shows the median and mean of the features and the resulting p-values of the t-tests. Only one feature shows a statistically significant difference: Wikidata items about women have a mean of 18.35 claims, while items about men have a mean of 19.74 claims ($p = 0.01$).

5.2.2 Associated Factors. Previous research in Wikipedia found that editors' attention and effort correlated strongly with the quality of Wikipedia articles. For example, the numbers of revisions to an article and the number of unique authors are strong predictors of article quality [28, 35]. This makes intuitive sense: more revisions indicates more effort, while more unique editors indicates more diverse perspectives.

Therefore, we want to see whether this relationship held in Wikidata: do more revisions and more unique editors for an item correlate with higher quality scores, as computed by ORES? The answer is yes. Using Spearman correlation, we found a strong positive association strong positive correlation between the number of revisions and ORES scores ($r_s(1602) = .78$, $p < .001$), and between the number of distinct editors and ORES score ($r_s(1602) = .73$, $p < .001$).

Finally, we check to see whether men and women Wikidata items differed in number of revisions and number of unique editors. Independent t-tests show that there is no significant difference in either case. For number of revisions, women items have a mean of 81.6 revisions (SD=54.5) and men items have a mean of 82.6 items (SD=52.0), ns ($p=0.762$). For number of unique editors, women items have a mean of 38.1 (SD = 21.9) and men items have a mean of 37.3 (SD=19.), ns ($p=0.523$).

6 DISCUSSION

Our guiding research question is: to what extent is Wikidata **reflecting real world gender bias** vs. **introducing additional gender bias**? Our analysis suggests answers.

We found that Wikidata editors are likely to over sample male-dominated professions such as American football and baseball, thus contributing to the general predominance of items representing men over items representing women. Our analysis that focused on a set of academic professions show that the gender distribution of Wikidata is no more biased than real world notability judgments in either coverage or quality. We next discuss some possible explanations for our results, and how the structured nature of Wikidata may lead to reduced bias. We also discuss some low quality Wikidata items we observed during our data collection process, which lets us discuss the role and importance of human effort. Finally, we mention the possible role of self-focus bias and identify directions for future work.

6.1 Wikidata's Factual Basis May Reduce Bias

One notable finding of our case study is that Wikidata's *coverage* of women vs. men is no more biased than real world notability assessments within a set of academic professions. The percentage of Wikidata items representing women in these professions is comparable to the percentage of women who received awards from the corresponding professional societies. More promising is the fact that the quality of items representing men and women is equivalent. This contrasts with studies of Wikipedia, which have shown biases in content about and relevant to women [26, 31, 37].

Several factors may explain why the quality of Wikidata items about men and women is comparable. First, Wikidata data for a person consists of facts about that person, such as name, date of birth, place of birth, country of citizenship, occupation, etc. More specifically, within a profession, additional properties might be prominent. For example, for politicians, these include their political party and elected position(s) held. Providing this sort of factual information about a person is more straightforward than editing a Wikipedia article. We conjecture that it does not offer as much opportunity for gender bias – even implicit types of bias – to creep

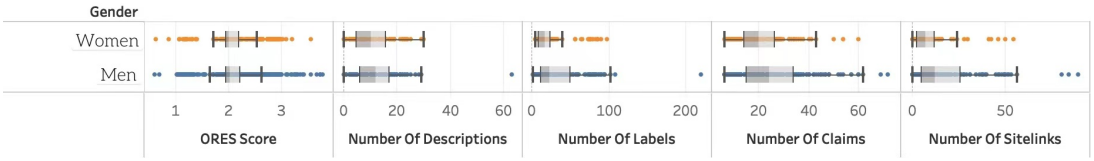


Figure 6: Box plots for ORES score and 4 features

Table 4: Summary of basic properties in men and women Wikidata pages. Statistically significant result is bolded.

	Women Wikidata Pages median / (mean)	Men Wikidata Pages median / (mean)	p-value
Number of Labels	11 (15.87)	11 (15.73)	0.887
Number of Descriptions	7 (8.78)	7 (9.07)	0.477
Number of Claims	**17 (18.35)**	**18 (19.74)**	**0.010***
Number of Sitelinks	2 (4.39)	2 (4.36)	0.962
ORES score	2.00 (2.11)	2.01 (2.13)	0.418

in, as has been shown in multiple comparisons of language used in Wikipedia biographies [2, 14, 38].

Further, much Wikidata content is added via automated bots, which import information from external data sources such as the Encyclopedia Brittanica. This also contributes to bringing in equivalent types and amounts of factual data about both women and men.

6.2 Human and bots both play vital roles

While bots are useful in bringing content into Wikidata, some exploratory analysis we did emphasizes the necessary roles of humans as well as bots. One of our filters for a *human* item to be included in our analysis is that it must have the properties gender, date of birth, country of citizenship and occupation. Without this information, it can be hard even to know which actual person an item refers to, and it obviously precludes many types of analysis. But 54% of human items did not pass this filter in our initial data collecting phase.

We encountered this problem when trying to determine whether people in our notability datasets are included in Wikidata. We sometimes found items with a matching name and perhaps a general profession such as *researcher* or *scientist* and one or two other properties, but we were not able to tell if this was the person in question. A bot might be able to import information about a person from a database, but a human Wikidata editor might be able to locate that person's website and to find and add additional important information. Future work could further explore the complementary role of human editors and bots and identify opportunities for tools to effectively combine human and automated effort.

6.3 Topical coverage and self-focus bias

A major source for the predominance of men items in Wikidata is the differential coverage of professions. Notably, three or four male-dominated sports professions are among the top five professions during each five-year interval. While it certainly is the case that some professions simply receive more attention, which makes them more likely to be covered in Wikidata, another reason may be playing a role: *self-focus bias*. Previous work has shown that contributors to peer production communities naturally enter and edit information on topics of interest to themselves [19]. Thus, *if* Wikidata editors consist mostly of men, *then* self-focus bias likely is contributing to this particular gender coverage gap. Future work to investigate the demographics of Wikidata editors would be helpful.

7 SUMMARY

We conducted a case study of the gender content gap in Wikidata. We began by noting that only 22% of Wikidata items representing people are about women. This led us to ask: was this due to existing real-world biases, or was it due to decisions of Wikidata editors? We answered this question by comparing Wikidata data to two external datasets, US Bureau of Labor Statistics data that showed the overall gender distribution within professions, and lists of award winners by a set of professional societies, which indicate who is considered "notable" within those professions. We found that Wikidata's representation of women within a set of professions was comparable to the professional societies' notability assessments, and both contained lower proportions of women than in the profession as a whole. We also observed that many of the professions with most items in Wikidata are male-dominated sports professions. Finally, we found that the quality of Wikidata items representing women was comparable to the quality of items representing men. We discussed several implications and possible next steps based on our findings.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their comments and suggestions that helped us strengthen our paper. This work was supported by the National Science Foundation(NSF) under Award No. IIS-1816348.

REFERENCES

[1] Judd Antin, Raymond Yee, Coye Cheshire, and Oded Nov. 2011. Gender differences in Wikipedia editing. In *Proceedings of the 7th international symposium on wikis*

and open collaboration. 11–14.
[2] David Bamman and Noah A Smith. 2014. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics* 2 (2014), 363–376.
[3] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1301–1310.
[4] Julia B Bear and Benjamin Collier. 2016. Where are the women in Wikipedia? Understanding the different psychological experiences of men and women in Wikipedia. *Sex Roles* 74, 5-6 (2016), 254–265.
[5] Benjamin Cabrera, Björn Ross, Marielle Dado, and Maritta Heisel. 2018. The Gender Gap in Wikipedia Talk Pages. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[6] Shuo Chang, Vikas Kumar, Eric Gilbert, and Loren G Terveen. 2014. Specialization, homophily, and gender in a social curation site: Findings from Pinterest. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 674–686.
[7] Benjamin Collier and Julia Bear. 2012. Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 383–392.
[8] Danielle J Corple. 2016. Beyond the Gender Gap: Understanding Women's Participation in Wikipedia. (2016).
[9] Maitraye Das, Brent Hecht, and Darren Gergle. 2019. The gendered geography of contributions to OpenStreetMap: Complexities in self-focus bias. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[10] Casey Fiesler, Shannon Morrison, and Amy S Bruckman. 2016. An archive of their own: a case study of feminist HCI and values in design. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2574–2585.
[11] Z Gardner, Peter Mooney, S De Sabbata, and L Dowthwaite. 2020. Quantifying gendered participation in OpenStreetMap: responding to theories of female (under) representation in crowdsourced mapping. *GeoJournal* 85, 6 (2020), 1603–1620.
[12] Rishab A Ghosh, Ruediger Glott, Bernhard Krieger, and Gregorio Robles. 2002. Free/libre and open source software: Survey and study.
[13] Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. 2010. Wikipedia survey–overview of results. *United Nations University: Collaborative Creativity Group* 8 (2010), 1158–1178.
[14] Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in Wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. 165–174.
[15] Aaron Halfaker. 2017. Interpolating quality dynamics in Wikipedia and demonstrating the Keilana effect. In *Proceedings of the 13th International Symposium on Open Collaboration*. 1–9.
[16] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.
[17] Aaron Halfaker, Jonathan Morgan, Amir Sarabadani, and Adam Wight. 2016. ORES: Facilitating re-mediation of Wikipedia's socio-technical problems. *Working Paper, Wikimedia Research* (2016).
[18] Eszter Hargittai and Aaron Shaw. 2015. Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia. *Information, communication & society* 18, 4 (2015), 424–442.
[19] Brent Hecht and Darren Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on communities and technologies*. 11–20.
[20] Benjamin Mako Hill and Aaron Shaw. 2013. The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS one* 8, 6 (2013), e65782.
[21] Laura Hollink, Astrid Van Aggelen, and Jacco Van Ossenbruggen. 2018. Using the web of data to study gender differences in online knowledge sources: the case of the European parliament. In *Proceedings of the 10th ACM Conference on Web Science*. 381–385.
[22] Daniela Iosub, David Laniado, Carlos Castillo, Mayo Fuster Morell, and Andreas Kaltenbrunner. 2014. Emotions under discussion: Gender, status and communication in online collaboration. *PloS one* 9, 8 (2014), e104880.
[23] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3819–3828.
[24] Maximilian Klein, Harsh Gupta, Vivek Rai, Piotr Konieczny, and Haiyi Zhu. 2016. Monitoring the gender gap with wikidata human gender indicators. In *Proceedings of the 12th International Symposium on Open Collaboration*. 1–9.
[25] Rebecca Knowles, Josh Carroll, and Mark Dredze. 2016. Demographer: Extremely simple name demographics. In *Proceedings of the First Workshop on NLP and Computational Social Science*. 108–113.
[26] Shyong (Tony) K Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R Musicant, Loren Terveen, and John Riedl. 2011. WP: clubhouse? An exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th international symposium on Wikis and open collaboration*. 1–10.
[27] David Laniado, Andreas Kaltenbrunner, Carlos Castillo, and Mayo Fuster Morell. 2012. Emotions and dialogue in a peer-production community: the case of Wikipedia. In *proceedings of the eighth annual international symposium on wikis and open collaboration*. 1–10.
[28] Andrew Lih. 2004. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *Nature* 3, 1 (2004), 1–31.
[29] Shlomit Aharoni Lir. 2019. Strangers in a seemingly open-to-all website: the gender bias in Wikipedia. *Equality, Diversity and Inclusion: An International Journal* (2019).
[30] Amanda Menking and Ingrid Erickson. 2015. The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 207–210.
[31] Amanda Menking, David W McDonald, and Mark Zachry. 2017. Who Wants to Read This? A Method for Measuring Topical Representativeness in User Generated Content Systems. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2068–2081.
[32] Amanda Menking and Jon Rosenberg. 2021. WP: NOT, WP: NPOV, and Other Stories Wikipedia Tells Us: A Feminist Critique of Wikipedia's Epistemology. *Science, Technology, & Human Values* 46, 3 (2021), 455–479.
[33] Joseph Reagle. 2013. "Free as in sexist?" Free culture and the gender gap. *first monday* (2013).
[34] Monica Stephens. 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal* 78, 6 (2013), 981–996.
[35] Besiki Stvilia, Michael B Twidale, Linda C Smith, and Les Gasser. 2005. Assessing Information Quality of a Community-Based Encyclopedia. *ICIQ* 5, 2005 (2005), 442–454.
[36] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. 2012. Gender, representation and online participation: A quantitative study of stackoverflow. In *2012 International Conference on Social Informatics*. IEEE, 332–338.
[37] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9.
[38] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science* 5 (2016), 1–24.
[39] Elijah Zolduoarrati and Sherlock A Licorish. 2021. On the Value of Encouraging Gender Tolerance and Inclusiveness in Software Engineering Communities. *Information and Software Technology* (2021), 106667.

A APPENDIX: 133 WIKIDATA PROFESSIONS WITH MORE THAN 100 USA DATA ITEMS

Wikidata Qid	Occupation	Number of items
Q19204627	American football player	16090
Q33999	actor	12706
Q82955	politician	8083
Q3665646	basketball player	8019
Q10871364	baseball player	7413
Q937857	association football player	4386
Q177220	singer	4148
Q36180	writer	3391
Q639669	musician	3258
Q1930187	journalist	2762
Q40348	lawyer	2603
Q11774891	ice hockey player	2352
Q2526255	film director	1979
Q36834	composer	1939
Q488205	singer-songwriter	1830
Q488111	pornographic actor	1815
Q483501	artist	1679
Q6625963	novelist	1658
Q28389	screen writer	1516
Q11513337	athletics competitor	1412
Q11338576	boxer	1120
Q43845	businessperson	1120
Q10798782	television actor	1053
Q2252262	rapper	1024
Q11303721	golfer	993
Q13382576	rower	957
Q33231	photographer	921
Q13474373	professional wrestler	919
Q1028181	painter	909
Q11607585	mixed martial artist	877
Q4610556	model	839
Q10833314	tennis player	813
Q2309784	sport cyclist	811
Q753110	songwriter	800
Q2066131	athlete	787
Q49757	poet	774
Q15981151	jazz musician	725
Q10800557	film actor	724
Q3282637	film producer	723
Q15117302	volleyball player	719
Q5137571	basketball coach	687
Q10843402	swimmer	667
Q81096	engineer	656
Q201788	historian	614
Q183945	record producer	608
Q855091	guitarist	603
Q2405480	voice actor	578
Q19595175	amateur wrestler	513
Q2722764	radio personality	512
Q386854	drummer	501
Q189290	military officer	474
Q378622	racing driver	474
Q131524	entrepreneur	454

Continued on next page

Table 5 – *Continued from previous page*

Wikidata Qid	Occupation	Number of items
Q82594	computer scientist	453
Q170790	mathematician	413
Q482980	author	409
Q13219587	figure skater	409
Q188094	economist	406
Q3246315	head coach	401
Q212980	psychologist	341
Q715301	comics artist	335
Q130857	disc jockey	324
Q19841381	Canadian football player	312
Q2259451	stage actor	302
Q3400985	academic	298
Q245068	comedian	280
Q1238570	political scientist	273
Q41583	coach	271
Q486748	pianist	264
Q169470	physicist	253
Q4009406	sprinter	246
Q4964182	philosopher	236
Q4773904	anthropologist	235
Q4144610	alpine skier	224
Q158852	conductor	220
Q10349745	racing automobile driver	217
Q947873	television presenter	216
Q2986228	sports commentator	213
Q13381572	artistic gymnast	212
Q1114448	cartoonist	207
Q17682262	lacrosse player	207
Q15295720	poker player	206
Q1281618	sculptor	206
Q11063	astronomer	195
Q47064	military personnel	194
Q37226	teacher	188
Q193391	diplomat	187
Q1622272	university teacher	180
Q578109	television producer	180
Q2961975	business executive	178
Q593644	chemist	168
Q864503	biologist	163
Q214917	playwright	156
Q42973	architect	154
Q584301	bassist	153
Q250867	Catholic priest	152
Q16533	judge	150
Q13381863	fencer	144
Q10873124	chess player	144
Q18581305	beauty pageant contestant	144
Q15709642	snowboarder	143
Q17502714	skateboarder	139
Q3014296	motorcycle racer	136
Q222344	cinematographer	136
Q14089670	rugby union player	135
Q39631	physician	134
Q2490358	choreographer	133

Continued on next page

Table 5 – *Continued from previous page*

Wikidata Qid	Occupation	Number of items
Q6665249	judoka	132
Q644687	illustrator	129
Q16029547	biathlete	127
Q846750	jockey	126
Q5716684	dancer	125
Q3501317	fashion designer	123
Q484876	chief executive officer	119
Q17524364	water polo player	117
Q901	scientist	117
Q13561328	surfer	115
Q11631	astronaut	113
Q17125263	YouTuber	112
Q13388586	softball player	110
Q2374149	botanist	109
Q10866633	speed skater	108
Q18617021	freestyle skier	108
Q2306091	sociologist	108
Q3499072	chef	108
Q15982858	motivational speaker	107
Q2059704	television director	106
Q484188	serial killer	104
Q15253558	activist	103
Q14467526	linguist	102
Q10843263	field hockey player	102
Q13141064	badminton player	100