Supplemental Digital Appendix 1

Expanded Method for Matching Each PLOS Article in the Data Set With its Corresponding WoS Record, From a Study of Author Contributorship Roles and Gender, 2008–2013

Method

Data sources

Two sources of data are used: Thomson Reuters' Web of Science (WoS) and all articles published by the Public Library of Science (PLOS), available on the PLOS website in XML format.

The WoS database includes the Science Citation Index Expanded (SCIe), the Social Science Citation Index (SSCI), and the Art & Humanities Citation Index (AHCI). As of 2014, the WoS covers more than 50 million articles published in almost 20,000 journals.

The Public Library of Science (PLOS) publishes 8 peer-reviewed scientific journals. PLOS Biology (2003) and PLOS Medicine (2004) were the two first journals founded, followed by PLOS Genetics, PLOS Computational Biology and PLOS Pathogens in 2005, by PLOS ONE in 2006 and PLOS Neglected Tropical Diseases in 2007. PLOS Clinical Trials was published between 2006 and 2007.

Between 2003 and 2014, 127,911 articles were published in PLOS journals (eTable 1). PLOS ONE is the most prolific, with 106,460 documents published between 2006 and October 9, 2014. Other documents (n=21,451) were published across the seven other journals.

eTable 1. Number of documents per journal, 2003–2014

Journal	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	All
PLOS Biology	98	456	431	423	321	327	264	304	276	230	292	196	3,618
PLOS Computational Biology			72	168	251	287	376	414	418	521	553	423	3,483
PLOS Clinical Trials				40	28								68
PLOS Genetics			77	208	230	352	473	471	565	721	874	552	4,523
PLOS Medicine		68	434	487	346	250	199	193	206	208	219	128	2,738
PLOS Neglected Tropical Diseases					42	179	224	350	445	525	623	533	2,921
PLOS ONE				137	1,230	2,716	4,405	6,750	13,797	23,464	31,524	22,437	106,460
PLOS Pathogens			41	123	198	286	459	534	556	640	739	524	4,100
All journals	98	524	1,055	1,586	2,646	4,397	6,400	9,016	16,263	26,309	34,824	24,793	127,911

Data processing

Download and extraction

In order to create the corpus on the contributions of each of the authors, we used the data compiled by PLOS within the framework of the Article-Level Metrics[1]. Available in an Excel file[2], these data provide various citation and usage indicators but, more importantly for this project, the DOI of each of the

[1] See : http://article-level-metrics.PLOS.org/alm-info/

[2] Available at : http://article-level-metrics.PLOS.org/PLOS-alm-data/.

articles, which was used to 1) download the full text of each PLOS article and 2) match each PLOS article with its record in the WoS.

In addition to full-text, PLOS journals make available PDF, RIS, BibTex and XML versions of the articles. In order to build the URLs of each of these articles, PLOS uses a standard format, which includes the journal URL, DOI, and type of format. For example, the link [http://www.PLOSone.org/article/fetchObjectAttachment.action?uri=info%3A**doi**%2F10.1371%2Fjournal.pone.0004048&representation=XML] retrieves the XML format of the article "The Effects of Aging on Researchers' Publication and Citation Patterns".

Using this structure, we built a code for automatically downloading the XML format of each PLOS paper. Using the DOIs found in the Article-Level-Metrics table, the URL of the XML format of each of the documents was built and queried using the SQL Server Integration Services (SSIS) as well as a Visual C# code (see eAppendix 1). This uploads automatically the XML format of each article to the user's computer (see eAppendix 2).

Given that the metadata of each PLOS paper is available on WoS, only two elements are needed from the PLOS articles' XML structure: the DOI and list of authors' contributions. DOIs are kept in order to match articles with the WoS. In order to isolate and retrieve the author's contribution from the full text of the articles, another Visual C# script integrated to SSIS was written. Articles' DOIs were obtained through their URLs.

Table 3 provides the number of PLOS articles retrieved and the number of PLOS articles in WoS, as well as the proportion of PLOS articles that were matched with the WoS. As shown, 97.6 % (n=94,879) of all PLOS articles were indexed in the WoS. However, given that not all PLOS articles were assigned a DOI in the WoS, there is a small fraction of PLOS articles that could not be matched to the WoS. On the whole, more than 95.5% of PLOS articles published between 2008 and 2013 were matched to the WoS (92,845). Most of the articles are published in the journal PLOS ONE (more than 85%), and the large majority of these could be matched with the WoS (98.2%). The journal PLOS Clinical Trials, which was only published between 2006 and 2007, was excluded from the analysis, as it did not published any articles during the period covered (eTable 2).

eTable 2. Number and percentage of papers published in PLOS journals indexed in the WoS, 2008 - 2013

Journal	PLoS	WoS	Direct link with DOI	Match (%)
PLOS Biology	1,693	1,380	988	58.4%
PLOS Computational Biology	2,569	2,429	2,090	81.4%
PLOS Genetics	3,456	3,251	2,865	82.9%
PLOS Medicine	1,275	1,214	975	76.5%
PLOS Neglected Tropical Diseases	2,346	2,197	2,098	89.4%
PLOS ONE	82,656	81,393	81,208	98.2%
PLOS Pathogens	3,214	3,015	2,621	81.5%
All PLOS Journals	97,209	94,879	92,845	95.5%

Selection of the corpus

The dataset of PLOS articles, including authors' contributions as well as all WoS metadata, served as the sampling frame for the study. From this, any document types that were not standard articles and review articles were excluded, given that these are more likely to represent original contributions to knowledge (Moed, 2006). Furthermore, only articles published between 2008 and 2013 were included, as WoS only provided full first names and links to institutional addresses for this period (this was fundamental in the gender-name assignment procedure). This reduced the total number of articles to 88,067. Also excluded were articles lacking author contribute data (n=962) as well as those for which a match could not be established between PLOS and WoS (n=369). Four duplicates were also excluded. The final dataset included 87,002 articles published in the seven PLOS journals between 2008 and 2013.

Parsing the contributions field

For each article, each PLOS journal provides a list of each author's contribution to predetermined categories of contributions. The most common contributions are:

- Analyzed the data
- Performed the experiments
- Conceived and designed the experiments
- Wrote the paper
- Contributed reagents/materials/analysis tools

However, these statements of contributorship can take several forms. The most common form is this one:

> Conceived and designed the experiments: SD RK. Performed the experiments: SD MM MJH. Analyzed the data: SD MM. Contributed reagents/materials/analysis tools: MJH. Wrote the paper: SD RK.

In this case, the name of the contribution is followed with a colon, a space, and then the initials of each of the authors who have performed this task. Each author is separated by a space. There are, however, other forms which are more difficult to parse. In the following example, all authors' initials are separated by a comma, and are followed by the specific contribution.

> MS, CS, NC, CT, FGB, DR, NSF, MCP, HF, MPF, FB, PVA, PEC, SO, AG, FAS, PD, AM, MLA, and OS conceived, designed or performed the experiments. MS, CS, NC, CT, FGB, DR, NSF, KLR, MCP, HF, FB, PVA, PEC, SO, AG, IS, FAR, FAS, PD, AM, MLA, and OS analyzed the data or corrected the paper. BV, AZ, and AS contributed reagents/materials/analysis tools. MS, MLA, and OS wrote the paper.

This one, on the other hand, has the names of the authors written at length.

> Conceived and designed the experiments: Ohad Yogev, Orli Yogev, Eitan Shaulian, Michal Goldberg, Thomas D. Fox, Ophry Pines. Performed the experiments: Ohad Yogev, Orli Yogev, Esti Singer, Thomas D. Fox. Analyzed the data: Ohad Yogev, Orli Yogev, Eitan Shaulian, Michal Goldberg, Thomas D. Fox, Ophry Pines. Contributed reagents/materials/analysis tools: Michal Goldberg, Thomas D. Fox, Ophry Pines. Wrote the paper: Ohad Yogev, Michal Goldberg, Ophry Pines.

Extraction of authors' contributions

The first step in extracting author contributions was to identify the most common form of contribution statements. In order to do so, we have limited the first step to the 86,725 statements that start by seven characters (without a space) or by the following words: *wrote, final, first, ICMJE, model, this, the, took, idea, for, gave, data, built, study.* The name of the contribution was then isolated, as well as the initials (or names) of the researchers who have performed them.

To reduce the problems associated with the various forms of author contributions, we have divided our dataset into two categories. The first group comprises all contribution statements where there are equal numbers of colon and end points (n=82,031) while the second group consists of those for which the number of colons and end points is not identical (n=4,694). This allows distinguishing the contributions where end points are used for functions other than the end of a phrase or where points are missing and, would, thus, compromise the parsing of initials in subsequent treatments.

For the first group, we start by dividing the text using the end point found at the end of each contribution. For example, the following statement:

> Conceived and designed the experiments: SD RK. Performed the experiments: SD MM MJH. Analyzed the data: SD MM. Contributed reagents/materials/analysis tools: MJH. Wrote the paper: SD RK.

Is divided into the following sentences:

1. Conceived and designed the experiments: SD RK
2. Performed the experiments: SD MM MJH
3. Analyzed the data: SD MM
4. Contributed reagents/materials/analysis tools: MJH
5. Wrote the paper: SD RK

These contribution statements are then separated into two sections using the colon: the left part, which contains the contribution, and the right part, which contains the initials of the authors who have performed the task. Each initial is also extracted and placed into a specific field.

For the other group of contribution statements where the number of colons and end points is not identical, we cannot use the point as the marker of the end of a contribution. For example, this contribution statement contains a point following one of the authors' initials (Thomas D. Fox):

> Conceived and designed the experiments: Ohad Yogev, Orli Yogev, Eitan Shaulian, Michal Goldberg, Thomas D. Fox, Ophry Pines. Performed the experiments: Ohad Yogev, Orli Yogev, Esti Singer, Thomas D. Fox. Analyzed the data: Ohad Yogev, Orli Yogev, Eitan Shaulian, Michal Goldberg, Thomas D. Fox, Ophry Pines. Contributed reagents/materials/analysis tools: Michal Goldberg, Thomas D. Fox, Ophry Pines. Wrote the paper: Ohad Yogev, Michal Goldberg, Ophry Pines.

Hence, for this type of contribution, we need to find an additional way of separating the various contributions. We thus introduced another marker in each of the statements to parse each of the contributions. A vertical bar ' | ' was thus inserted at the beginning of each contribution to replace the point as a marker of contributions. The colon was then used to isolate the initials from the rest of the

contribution field, and then, using the vertical bar, we isolated the contribution label. Here's an example of this treatment:

> |conceived and designed the experiments: ohad yogev, orli yogev, eitan shaulian, michal goldberg, thomas d. fox, ophry pines. |performed the experiments: ohad yogev, orli yogev, esti singer, thomas d. fox. |analyzed the data: ohad yogev, orli yogev, eitan shaulian, michal goldberg, thomas d. fox, ophry pines. |contributed reagents/materials/analysis tools: michal goldberg, thomas d. fox, ophry pines. |wrote the paper: ohad yogev, michal goldberg, ophry pines.

Then, in a manner similar to the first group, we have isolated each contributing author using the space or comma. After a few treatments on the contribution statements that did not follow exactly the form typically used, we obtained collaboration statements for all 87,002 articles. On the whole, we obtained 20,667 distinct contribution labels, associated with 40,356 initials (contributors), for more than 1.5 million records. After the cleaning of contribution statements with typos, as well as the grouping of contribution statements having the same signification (for example, 'writing the paper' and 'writing the manuscript'), we obtained a list of the most common contributions, as well as the number of articles and of author-article combinations that feature this contribution (eTable 3).

eTable 3. Number and percentage of articles and of article-initial pairs, by contribution label

	Author-initial combinations		Articles	
Contribution	**N**	**%**	**N**	**%**
Analyzed the data	320,080	50.6%	85,900	98.7%
Performed the experiments	311,679	49.3%	82,811	95.2%
Conceived and designed the experiments	288,765	45.6%	85,406	98.2%
Wrote the paper	287,796	45.5%	86,517	99.4%
Contributed reagents/materials/analysis tools	220,331	34.8%	64,444	74.1%
Other (20 243)	*79,978*	*12.6%*	*15,900*	*18.3%*
N. distinct papers	**632,799**	**-**	**87,002**	**-**

Establishing a link between the WoS and the PLOS database

As mentioned previously, the DOI was used to match PLOS contributions and bibliographic information found in the WoS. Another link also has to be established between the two data sources, as the initials of the authors from the PLOS database have to be linked with the names of the authors found in the WoS record, in order to obtain each authors' gender. Each authors' name from the WoS was thus transformed into an 'initials' format to match the PLOS contributions using a SQL script. For more than 98% of articles (n=85,260), all authors were assigned to their contribution (eTable 4). Those that were not matched were due to mistakes in the spelling of the names in WoS, or their absence in the authors' list.

eTable 4. Number and percentage of articles and of article-initial pairs matched with the WoS, by contribution label

	Articles not matched with WoS		Articles matched with WoS		Percent difference	
Contribution	N article-initial pairs	N articles	N article-initial pairs	N articles	N article-initial pairs	N articles
Analyzed the data	320,080	85,900	306,592	84,221	4.2%	2.0%
Performed the experiments	311,679	82,811	297,893	81,183	4.4%	2.0%
Conceived and designed the experiments	288,765	85,406	277,302	83,734	4.0%	2.0%
Wrote the paper	287,796	86,517	274,615	84,789	4.6%	2.0%
Contributed reagents/materials/analysis tools	220,331	64,444	208,794	63,049	5.2%	2.2%
*Other**	*79,978*	*15,900*	*67,929*	*15,416*	15.1%	*3.0%*
N. distinct papers	**632,799**	**87,002**	**589,892**	**85,260**	**6.8%**	**2.0%**

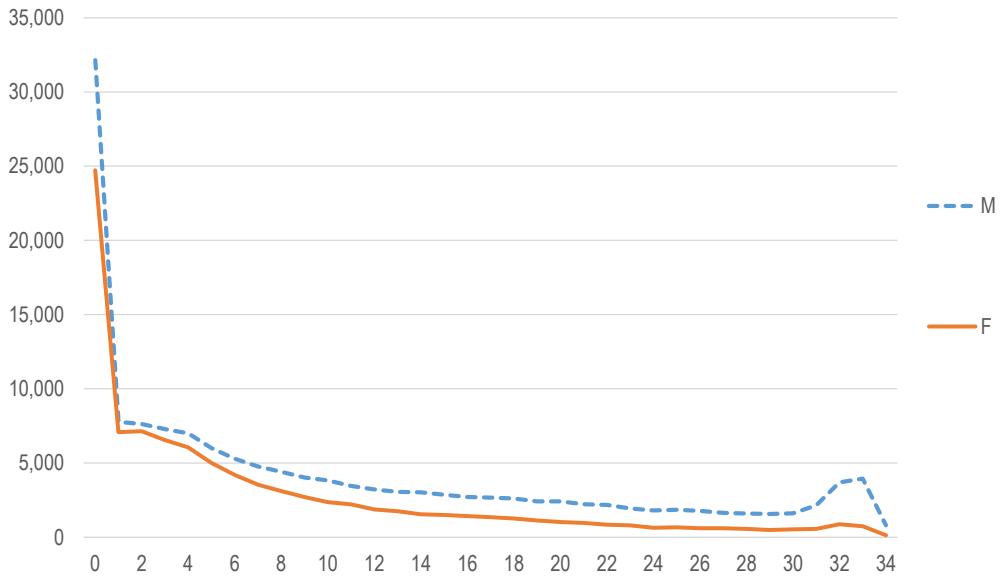
*Including the original wording of "Other" contributions as well as all other categories that did not correspond with these five main categories

Gender and age assignation

The gender of authors has been attributed using gender assignation tables developed in Larivière et al. (2013). This list uses given names and country combinations to assign gender to authors of articles. On the whole, this conversion list managed to assign a gender to 88.1% of authorships – i.e., author-paper combinations – found in the paper, of which 32.5% were female and 55.7% were male. Initials and unisex names accounted for 0.2% and 2.7% respectively, while unknown was 8.9%. This unknown rate is slightly above that obtained by Larivière et al. (2013) for all WoS (8.4%).

Academic age of authors was estimated using their year of first publication, as recorded in the Web of Science. In order to obtain such age, authors found in the WoS were disambiguated automatically by the Center for Science and Technology Studies (CWTS, Leiden University) using the algorithm developed by Caron& van Eck (2014). The distribution of ages for male and female contributorships is presented in eFigure 1. Given that the majority (95%) of authors were (academically) younger than 30 years—and that ages above 30 are overestimated because the first "possible" publication year in our dataset is 1980, the analyses for this paper focus on those whose academic age is between 0 and 30 years.

eFigure 1. Number of female and male contributorships by academic age.



Regression analysis

An analytic sample was constructed for the regression using those observations that contained all necessary variables (i.e., author position, academic age, number of authors, percent of female authors, country, and discipline): 270,103 observations were used of 589,906 possible observations. eTables 5 and 6 detail the differences between these samples. Regressions were run on the entire dataset and using each contributorship type as a dependent variable. The unit of analysis was author-publication combinations.

eTable 5. Differences between the whole and the regression samples

	Whole sample	Regression sample
n	589,906	270,103
Gender		
Females	32.5	32.4
Males	55.7	56.5
Unknown	11.9	11.1
First author	14.5	14.6
Last author	14.5	14.4
Corresponding author	14.4	14.6
Field		
Arts	.0	.0
Biology	5.6	5.0
Biomedical Research	43.2	46.5
Chemistry	.3	.2
Clinical Medicine	47.8	45.8
Earth and Space	.6	.4
Engineering and Tech	.3	.2
Health	.5	.4
Humanities	.0	.0
Mathematics	.1	.1
Physics	.3	.3
Professional Fields	.1	.1
Psychology	1.0	.9
Social Sciences	.2	.2
Unknown	.0	.0
Contributions		
analysis	52.0	52.8
design	47.0	47.0
material	35.4	35.0
perform	50.5	50.5
write	46.6	46.3
Publication year		
2008	3.1	5.9
2009	2.8	5.2
2010	9.6	18.1
2011	17.9	33.9
2012	29.4	34.9
2013	37.3	2.0
Academic age[a]	11.0 (10.3)	11.0 (10.3)
Number of authors[a]	6.9 (4.2)	7.0 (4.4)
Percent female[a]	35.6 (24.9)	35.3 (24.6)

Notes: Values in table show percentage of observations

[a]Descriptive analysis at document level. Numbers in corresponding rows are mean and standard deviation in parentheses. n=85,260 in whole sample and 49,290 in regression sample

eTable 6. Difference between whole and regression samples for countries with more than 0.5% of whole sample

	Whole sample	Regression sample
n	589,906	270,103
number of countries	189	178
missing	1.2	0
AUSTRALIA	2.8	2.8
AUSTRIA	.7	.6
BELGIUM	1.0	1.1
BRAZIL	1.7	1.5
CANADA	3.1	3.2
DENMARK	.9	.9
ENGLAND	5.1	5.7
FINLAND	.7	.7
FRANCE	5.3	5.9
GERMANY	6.3	6.5
INDIA	1.3	1.2
ISRAEL	.7	.8
ITALY	3.5	3.4
JAPAN	4.6	4.2
NETHERLANDS	2.5	2.6
NORWAY	.6	.6
PEOPLES-R-CHINA	12.4	9.1
PORTUGAL	.5	.5
SCOTLAND	.8	.8
SINGAPORE	.6	.6
SOUTH-KOREA	1.4	1.1
SPAIN	2.7	2.8
SWEDEN	1.7	1.8
SWITZERLAND	1.6	1.7
TAIWAN	1.8	1.4
USA	27.5	31.2

Supplemental Cited References

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In 19th International Conference on Science and Technology Indicators."Context counts: Pathways to master big data and little data" (pp. 79-86). CWTS-Leiden University Leiden.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C.R. (2013). Global gender disparities in science. Nature, 504(7479), 211-213.

eAppendix 1. SQL code for the creation of the URL of each PLOS paper

```
    SELECT DISTINCT idPLOS, SUBSTRING(DOI, CHARINDEX('/', doi) + 1, LEN(DOI)) AS doi,
SUBSTRING(URL, 1, CHARINDEX('.org', url) + 4) +
'article/fetchObjectAttachment.action?uri=info%3Adoi%2F10.1371%2F' +
SUBSTRING(DOI,CHARINDEX('/',doi) + 1,LEN(DOI)) + '&representation=XML'AS url
    FROM dbo.PLOS
    WHERE idPLOS = ? /*parameter*/
```

eAppendix 2. C# code for the download of each paper (through Microsoft SQL server Integration Services)

```
using System;
using System.IO;
using System.Xml;
using System.Data;
using Microsoft.SqlServer.Dts.Pipeline.Wrapper;
using Microsoft.SqlServer.Dts.Runtime.Wrapper;

[Microsoft.SqlServer.Dts.Pipeline.SSISScriptComponentEntryPointAttribute]

public class ScriptMain : UserComponent
{
 string fichiersource = string.Empty;
 string JournalTitle = string.Empty;
 string fntype = string.Empty;
 string contribution = string.Empty;

 public override void CreateNewOutputRows()
 {

  fichiersource = Variables.destination;
  XmlTextReader xmt = new XmlTextReader(fichiersource);

  while (xmt.Read())
  {

   if ((xmt.Name == "fn") && (xmt.HasAttributes))
   {
    xmt.MoveToAttribute("fn-type");
    fntype = xmt.GetAttribute("fn-type");

    if (fntype == "conflict")
    {
     xmt.MoveToAttribute("fn-type");
     fntype = xmt.GetAttribute("fn-type");
     if (fntype == "con")
     {
```

```
        xmt.ReadToFollowing("p");
        contribution = xmt.ReadInnerXml();

        this.Output0Buffer.AddRow();
        this.Output0Buffer.NomFichier = Variables.NomFichier;
        this.Output0Buffer.DOI = "";
        this.Output0Buffer.Contribution = contribution;
        xmt.Close();
       }
      }
      else if (fntype == "con")
      {
       xmt.ReadToFollowing("p");
       contribution = xmt.ReadInnerXml();

       this.Output0Buffer.AddRow();
       this.Output0Buffer.NomFichier = Variables.NomFichier;
       this.Output0Buffer.DOI = "";
       this.Output0Buffer.Contribution = contribution;
       xmt.Close();
      }
      }

    this.Output0Buffer.EndOfRowset();
   }
  }
}
```

Appendix 3. C# code for extraction author's contribution in PLOS articles

```
    using System;
    using System.IO;
    using System.Xml;
    using System.Data;
    using Microsoft.SqlServer.Dts.Pipeline.Wrapper;
    using Microsoft.SqlServer.Dts.Runtime.Wrapper;

    [Microsoft.SqlServer.Dts.Pipeline.SSISScriptComponentEntryPointAttribute]

    public class ScriptMain : UserComponent
    {
     string fichiersource = string.Empty;
     string JournalTitle = string.Empty;
     string fntype = string.Empty;
     string contribution = string.Empty;
```



```
public override void CreateNewOutputRows()
 {
  fichiersource = Variables.destination;
  XmlTextReader xmt = new XmlTextReader(fichiersource);

  while (xmt.Read())
  {

   if ((xmt.Name == "fn") && (xmt.HasAttributes))
   {
    xmt.MoveToAttribute("fn-type");
    fntype = xmt.GetAttribute("fn-type");

    if (fntype == "conflict")
    {
     xmt.MoveToAttribute("fn-type");
     fntype = xmt.GetAttribute("fn-type");
     if (fntype == "con")
     {
      xmt.ReadToFollowing("p");
      contribution = xmt.ReadInnerXml();

      this.Output0Buffer.AddRow();
      this.Output0Buffer.NomFichier = Variables.NomFichier;
      this.Output0Buffer.DOI = DOI;
      this.Output0Buffer.Contribution = contribution;
      xmt.Close();
     }
    }
   }
   else if (fntype == "con")
   {
    xmt.ReadToFollowing("p");
    contribution = xmt.ReadInnerXml();

    this.Output0Buffer.AddRow();
    this.Output0Buffer.NomFichier = Variables.NomFichier;
    this.Output0Buffer.DOI = DOI;
    this.Output0Buffer.Contribution = contribution;
    xmt.Close();

   }

   this.Output0Buffer.EndOfRowset();
  }
 }
}
```

