25 years of Social Work Research

Summary of Data Cleaning Procedures

Data Revolution

It is estimated that 90% of all the data in human history has been generated in the past two years, and the amount of data continues to grow by roughly 2.5 quintillion bytes every day (IBM, nd-a). We have unprecedented opportunities for deriving actionable insights from these data to advance the human condition. The significance continues to gain the attention of national and international organizations. For example, the Independent Expert Advisory Group (IEAG) of the United Nations prepared a report entitled, *A World That Counts: Mobilising the Data Revolution for Sustainable Development* (IEAG, 2014). The IEAG (2014) argues that "Data are the lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time; designing, monitoring and evaluating effective policies becomes almost impossible" (p. 2). The United States federal government established an open data initiative, motivated by the awareness that data can have the greatest impact when it is "accessible, discoverable, and usable" (www.data.gov). Data from all levels of government that is made completely open can lead to "cost savings, efficiency, fuel for business, improved civic services, informed policy, performance planning, research and scientific discoveries, transparency and accountability, and increased public participation in the democratic dialogue" (www.data.gov). At the time of preparing this article, over 124,000 data sets were freely available through www.data.gov among a diverse set of topic areas: agriculture, business, climate, consumers, ecosystems, education, energy, finance, health, local government, manufacturing, ocean studies, public safety, and science and research.

In 2012, the NIH established the Big Data to Knowledge (BD2K) initiative, recognizing the potential of data to advance our understanding of human health and disease. This initiative was designed to support four broad goals: 1) to facilitate use of biomedical digital assets by making them discoverable, accessible, and citable; 2) to conduct research and develop methods, software and tools to analyze biomedical big data; 3) to enhance training in the development of use and tools necessary for biomedical big data science; and 4) to support a data ecosystem that accelerates discovery as part of the digital enterprise (NIH). At roughly the same time, the National Science Foundation (NSF) created new initiatives to harness the potential of the data revolution, arguing that "data are motivating a profound transformation in the culture and conduct of scientific research in every field of science and engineering ... American scientists must rise to the challenges and seize the opportunities afforded by this new, data-driven revolution" (Suresh, XXX). The programs of the NSF are organized around capacity building and innovative applications. Capacity building activities focus on developing fundamental theories, techniques, methodologies and technologies that are broadly applicable to big data problems. Innovative applications relate to specific domains advancing innovative ideas that provide solutions "with potential for a broader impact on data science and its applications."

Data Science and Social Work

We are indeed living in a data revolution. The relevance to social work research is clear — we have the opportunity to study human behavior and other social phenomena in real time, with vastly improved reliability in our measurements, and at scale. Data comes from a wide range of sources, including (but not limited to) social media interactions, electronic medical records, audio and video recordings, click stream data, transaction data, web logs, embedded sensors, and GPS. New tools allow blending of data sets, unique ways to analyze data, and methods of visualizations that can produce insights limited only by one's imagination. As stated by the NIH, "lack of appropriate tools, poor data accessibility, and insufficient training, are major impediments to rapid translational impact" in this new era of data. It is essential that the field of social work fully embraces the data revolution and considers the extent to which the existing research infrastructure and training is preparing both social work researchers and providers to be effective and efficient in a data rich world. The potential of the data revolution is reflected in the current *Grand Challenge* entitled, "Harness

technology for social good." However, we believe that the data revolution goes beyond this single *Grand Challenge* as it provides the opportunity to be informative in all areas of social work research.

Data science serves as a useful framework for broadening the existing research infrastructure and training opportunities within social work. Data science is similar to traditional academic research, as it includes the key roles of substantive expertise and traditional research methodologies. Much of the theoretical basis comes from the field of statistics (Zumel & Mount, 2014), but the famous statistician William Cleveland argued that it is an interdisciplinary field that is much larger than statistics (Cleveland, XXX). A major difference is that data science draws heavily from knowledge in the areas of computer science, software engineering, and information technology. Unlike traditional academic research, data science is particularly well suited for tackling the *3 V's* of big data: *volume*, *variety*, and *velocity*. More specifically, in data-rich environments, researchers encounter an amount or volume of data that often exceeds the capacity of traditional database systems. Data also come in a variety of formats, ranging from highly structured data contained in relational databases to massive stores of unstructured text-data culled from social media. Many data opportunities go beyond the traditional approach of downloading a dataset and analyzing it with an off-the-shelf statistical package like SPSS, Stata, or SAS. Now, data can be accessed in real time through application program interfaces (API), which brings along the challenge of *velocity* — that is, data can be transmitted at a rate that grossly exceeds the computing power of desktop computers. These are practical challenges that need to be addressed in order to maximize the potential of these new data.

Another feature that distinguishes data science from traditional academic research is the focus on generating *actionable insights* from data rather than theory building and hypothesis testing. In fact, data science is more interested in deriving useful insights from correlational patterns as opposed to uncovering causal mechanisms. The focus is on the development of data products (Davenport, XXX), as opposed to reporting analysis. Data products are tools or services that generate actionable insights from the data itself, including (but not limited to) automated reports or dashboards, recommendation systems, prediction algorithms, decision tools, and interactive visualizations that help people with non-technical background make sense of complex data. Data science also values the philosophy of *openness* or open science. Thus, much effort is devoted to creating sustainable systems that make data as widely and freely accessible as possible, and developing and promoting the use of open source tools for managing and analyzing data. The concept of reproducible research is a *rule* rather than the *exception*. This involves making both data and statistical code available alongside research reports. This helps ensure the quality of the research, in addition to allowing researchers to make advances without having duplicate work that has already been done. Data science is by no means incompatible with traditional academic research. We believe that the tools, research strategies, and principles of data science are complementary to traditional ways of knowledge building in social work, offering many unique opportunities to improve the quality, efficiency, and impact of this area of research.

Overview of Current Study

In the current study, we seek to utilize the opportunities of the data revolution and data science by applying a range of tools and strategies to construct an historical base of social work research. In doing so we take on two of the 3 Vs of big data: volume due the amount of data collected and variety in terms of the multiple sources of that data. The data are article records harvested from existing bibliographic databases hosted on ProQuest and EbscoHost. Article records are already widely used in various bibliometric studies, particularly those that are intended to reveal publication networks within scientific communities. It does so by creating links between co-authors, and then visually representing the edges and vertices of the network. This is also done with article citation histories, to show influence over time. In this study, we were interested in the various article meta-data that are common to many scientific article records. For example, a single article record in psycINFO can have as many as XX pieces of meta-data including: dates, journal titles, article titles, author names, author affiliations, full text abstracts, keywords, subject classification, methodology, location, populations, page numbers, and source of funding. This is a rather large amount of information about a single paper, at least compared to the amount of information contained in article records before the availability of electronic storage and access. Large collections of article records may contain potentially interesting trends in the research, assuming the selection of the article records were conducted in a way that allowed meaningful inferences to be derived.

In the current study, we sought to capture every journal article record published in every social work journals over the in past quarter century (1989-2013) and to explore the various meta-data to uncover potentially interesting trends in the social work scholarship. For example, we can reasonably infer estimates of the *size* of social work scholarship by counting the number of unique article records. And, growth of the field can be inferred by examining the number of article records over time. Similarly, we examine the extent to which *team science* has become part of the research practice, as this has implications for training of social work researchers. We also use location meta-data to understand the extent to which different areas of the world have and have not been the focus of social work research. Finally, we used topic modeling to extract topic areas that define the focus of social work research. The execution of this research was grounded in the tools and values of data science. Thus, all data were managed and analyzed using the open source software, R. We also used an open source authoring system that integrates the statistical analysis with the text of the manuscript. Along with this manuscript, the authors will also release the actual data and statistical code, thereby meeting the requirements of *reproducible research*, as defined by King (). To our knowledge, this is the first time fully reproducible research (as defined by King, XXXX), has been published in a social work journal.

Davenport: http://blogs.wsj.com/cio/2014/06/25/so-you-want-to-build-a-data-product/

http://www.nsf.gov/pubs/2015/nsf15544/nsf15544.htm

http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607&org=NSF&from=news

Methods

Extraction of article records

Our list of social work journals was guided by the work of Hodge and Lacasse (2011). More specifically, Hodge and Lacasse (2011) identified 84 disciplinary social work journals based on a variety of sources, including *An Author's Guide to Social Work Journals* (NASW Press, 1997), Thyer's (2005) more recent listing of social work periodicals, and Genamics JournalSeek (http://journalseek.net/). Hodge and Lacasse (2011) examined the mission and aims of each journal, and eliminated journals that were specific to another field or had an had an inter-disciplinary focus. For the current study, our primary search query included all journal titles in this list.

To help ensure comprehensive coverage of all possible core social work journals, we created a supplemental search query that extracted journal titles with any of the following terms:

- "social work"
- "social welfare"
- "social casework"
- "social service"
- "human service"
- "social development"
- "social environment"

Using this search queries, we extract journal article recrods from three major databases on the EbscoHost platform: PsycINFO, Social Service Abstracts, and Social Work Abstracts. Because of known gaps in indexing in databases (Holden, Barker, Kuppens, Rosenberg & LeBreton, 2014; Holden, Barker, Covert-Vail, Rosenberg, & Cohen, 2009), we also searched ProQuest, which linked to an additional 46 minor databases, including (but not limited to): ERIC, Sociological Abstracts and Worldwide Political Science Abstracts. In every search, we used a filter in attempt to extract only article records that were classified as *journal articles*, thereby excluding other forms of scholarly communication that was not relevant to the current study (e.g., book reviews, editorials, obituaries, etc.). We limited the timeframe to a 25 year period, from 1989 to 2013

(inclusive). Articles published in years 2014 and 2015 were excluded because of delays in indexing. Search results were exported in batches of article records based on the restrictions of the platform. Article records were text files in a *generic bibliographic format*. The article records contained various meta-data based on the database from which the article record was extracted. Meta-data include (but are not limited to): article title, journal title, publication year, author name(s), author affiliation(s), abstract, keywords, methodological classification, funding source, location of study, subject groups, digital object identifier (DOI), number of references, number of pages, etc. These files were post-processed into a structured database using a set of scripts written in the R statistical programming language. The initial search resulted in 36,094 article records from 117 different journals. `NOTE: these values are hardcoded--should be changed to draw from actual data.`

The search queries were purposefully specified to be overly inclusive to ensure full coverage of all possible social work journals contained in the major and minor databases. Additionally, a visual inspection of the article records revealed problems in the original indexing of articles that requires subsequent data cleaning. Thus, our data cleaning procedures focused separately on journal titles and article records.

Data cleaning: Journal titles

The first step of data cleaning involved fixing journal title names due to discrepant errors in indexing. For example, *Journal of Gay and Lesbian Social Services* was indexed as a journal separate from *Journal of Gay & Lesbian Social Services* (use of "and" vs. "&"). Other examples involved journals indexed with and without subtitles, or journals with and without the word "The" at the beginning of the title. Additionally, some journals changed titles over their history. For example, *Journal of Technnology in Human Services* is formerly known as *Computers in Human Services*. In these situations, the former titles were merged with the current titles.

Many journals from allied health disciplines used one of the supplemental search terms in special editions, which were also included in the journal title. Thus, many non-social work journals were captured in the extraction process. To resolve this issue, we created a list of all journal titles that were not part of the core list defined by Hodge and Lacasse (2011). Study authors reviewed these journal titles and discussed whether each candidate title should be retained or excluded. When disagreements occurred, the study authors reviewed the mission and aims of the journals, names of editorial board members, and focus of the articles. A consensus was reached on all journal titles to be excluded and retained. After these procedures, the number of article records and journal titles was reduced to 34,956 and 82 (respectively).

Data cleaning: Article records

Our search procedures involved the use of three major databased and 46 minor databases, which resulted in some duplication of article records. Thus, a matching algorithm was constructed to identify and remove all duplicate article records. As previously noted, the filtering of only journal articles was not successful due to errors in the original indexing. Thus, other scholarly communications were captured in the extractio process. These other scholarly communications were readily identified due to specific patterns in the titles, such as the terms "Book Review," "From the Editor," and "Obituary." Given the number of article records in the database, it was not feasbile to manually extract these recrods. Instead, we created a separate database with representative examples of article records that we wanted to extract. We also added representative examples of article records to be retained. We then wrote a series of *regular expressions* to extract the problematic article records. Regular expressions are a sequence of characters that form search patterns, allowing extraction of records that match the specified pattern. In other words, it is a more sophisticated implementation of the search function common to all major word processors. We tested our regular expressions on the test database, achieving $> 95\%$ accuracy with regard to retaining and extracting article records. These regular expressions were then applied to the full database. We then excluded article records if they were missing essential meta-data, including author name(s), journal title, article title, and publication field. As an additional requirement, we required all articles to be ≥ 3 pages in length. Finally, we eliminated all social work journals with total article counts of < 10 for the enitre 25 year window (*Issues in Social Work Education*, *Maatskaplike*

Werk/Social Work, *Pediatric Social Work*, and *Critical Social Work*). These cleaning procedures resulted in a final database of 32,008 articles and 79 journals.

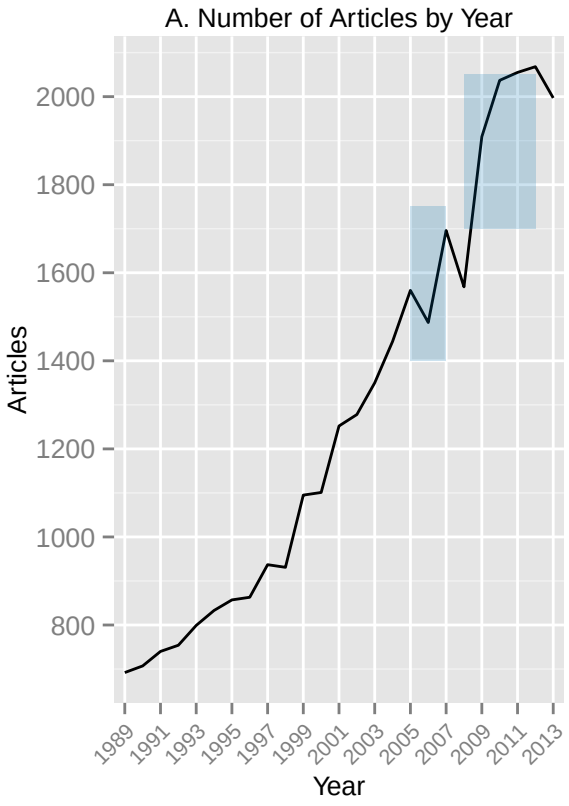
Data quality checks

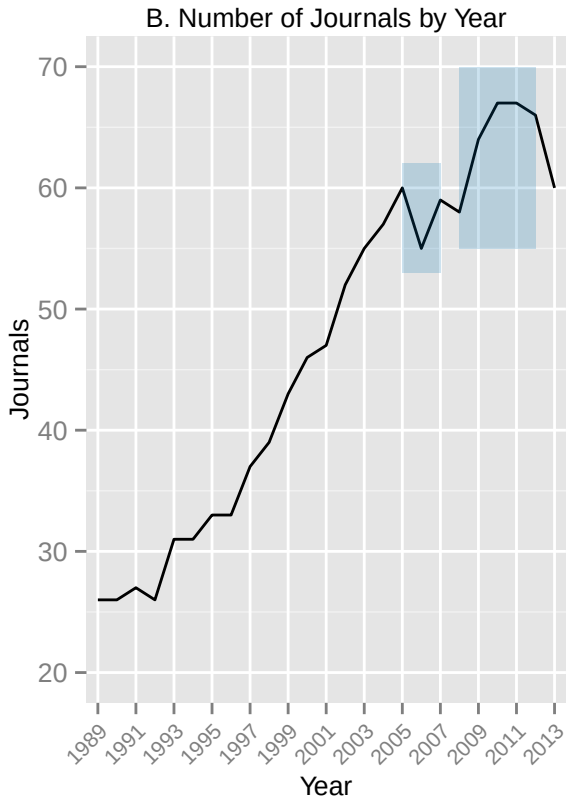
We performed a variety of checks on the final database. Because the size of the database, it was not feasible to manually inspect all records. Thus, we randomly sampled article records from the database and cross-checked the data we procured with the information listed with the journal's homepage. We also inspected a scatter plot of the number of article records for each journal by the number of years the journal appeared within the 25 window timeframe. Journals that exhibited significant deviation from the regression line were inspected to ensure that our cleaning procedures did not systematically exclude article records or retain other scholarly communications. All observed discrepancies were verified as differences in journals actual publication output. Finally, we also made checks of number of journal article records for various journals indexed in the *Web of Science*. Although the counts were close – i.e., within 10% – similar indexing problems observed from our record extraction were also observed in the *Web of Science* database, such as misclassification of book reviews as journal articles. Thus, we are unable to quantify the actual amount of error in our database is due to problems in indexing. At the same time, we are relying on the same source of data that social work researchers use to inform their work. Too much cleaning and post-processing of the search results can give rise to validity issues because social work researchers are using an information source that contain the errors we have attempted to eliminate. The issue of reliability and validity are given further attention in the discussion section of this study.

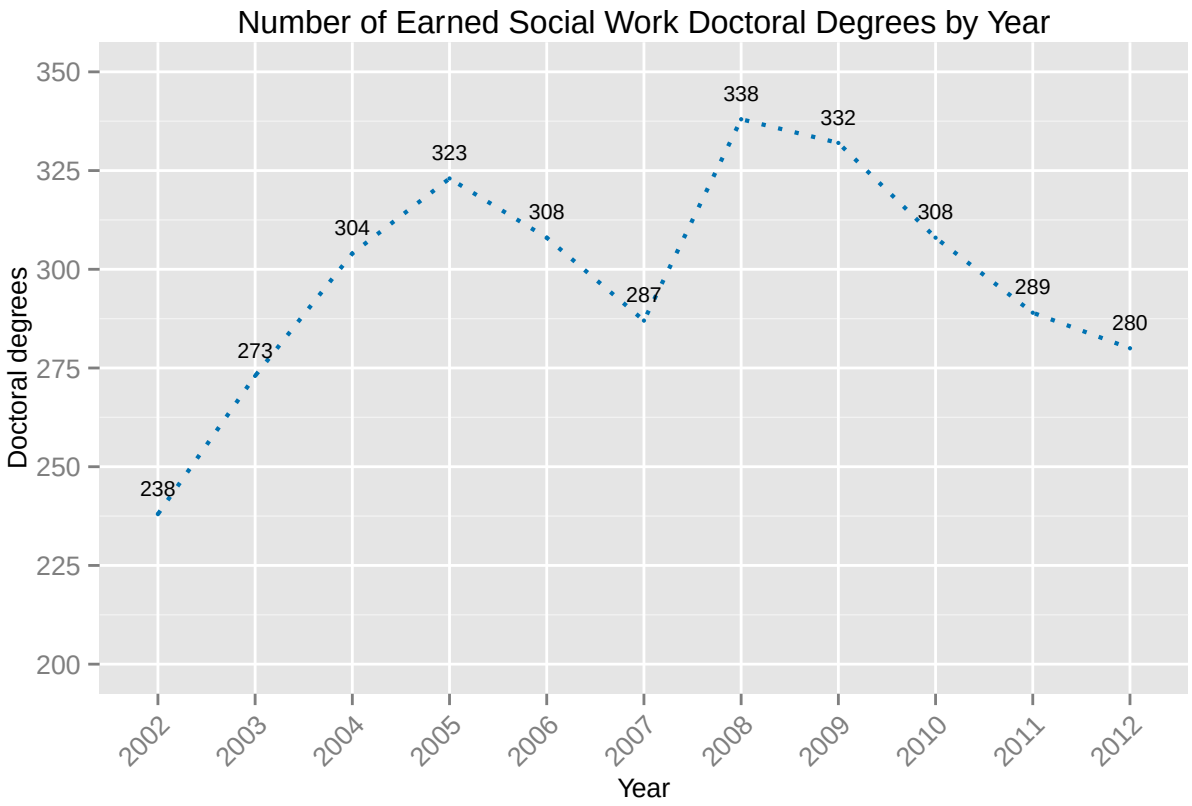
Results

title	first	last	n	H
Families in Society	1989	2013	1697	Y
Social Work	1989	2013	1457	Y
British Journal of Social Work	1989	2013	1413	Y
Journal of Gerontological Social Work	1989	2013	1271	Y
Research on Social Work Practice	1991	2013	1128	Y
Social Work in Health Care	1989	2013	1055	Y
International Social Work	1989	2013	1023	Y
Health & Social Work	1989	2013	947	Y
Journal of Social Work Education	1990	2013	900	Y
Journal of Sociology & Social Welfare	1989	2013	873	Y
Social Work Education	1999	2013	851	Y
Clinical Social Work Journal	1989	2013	807	Y
Journal of Human Behavior in the Social Environment	1997	2013	804	Y
Social Service Review	1989	2013	778	Y
Child & Adolescent Social Work Journal	1989	2013	744	Y
Indian Journal of Social Work	1989	2012	730	Y
Social Development	1992	2013	691	N
Social Work With Groups	1989	2013	663	Y
Reflections: Narratives of Professional Helping	1995	2013	655	Y
Journal of Social Service Research	1989	2013	639	Y
Journal of Psychosocial Oncology	1989	2013	600	Y
Administration in Social Work	1989	2013	590	Y
Child & Family Social Work	1997	2013	568	Y
Social Work Research	1994	2013	530	Y
Journal of Gay & Lesbian Social Services	1994	2013	511	Y
Smith College Studies in Social Work	1989	2013	489	Y
Journal of Technology in Human Services	1989	2013	445	Y
European Journal of Social Work	2000	2013	416	Y
Journal of Teaching in Social Work	1989	2013	414	Y
Canadian Social Work Review	1989	2013	392	Y
International Journal of Social Welfare	2000	2013	383	Y
Journal of Social Work Practice	1999	2013	382	Y
Arete	1989	2013	349	Y
Journal of Social Work Practice in the Addictions	2001	2013	347	Y
Social Work in Mental Health	2002	2013	298	Y
Journal of Baccalaureate Social Work	1995	2013	293	Y
Qualitative Social Work	2002	2013	281	Y
Journal of Ethnic & Cultural Diversity in Social Work	2000	2013	262	Y
Psychoanalytic Social Work	1989	2013	257	Y
Journal of HIV/AIDS & Social Services	2002	2013	254	Y
Journal of Social Work	2001	2013	254	Y
Australian Social Work	2006	2013	250	Y
Social Development Issues	1992	2013	245	Y
Advances in Social Work	2000	2012	235	Y
Children & Schools	2003	2013	230	Y
Journal of Family Social Work	1995	2013	220	Y
Social Work & Social Sciences Review	1989	2013	217	Y
Social Work & Society	2003	2012	211	Y
Professional Development	1998	2012	208	Y
Social Work & Christianity	2005	2013	186	Y
Social Work Review	2010	2013	185	Y
Affilia	2007	2013	184	Y
Journal of Religion & Spirituality in Social Work	2004	2013	177	Y

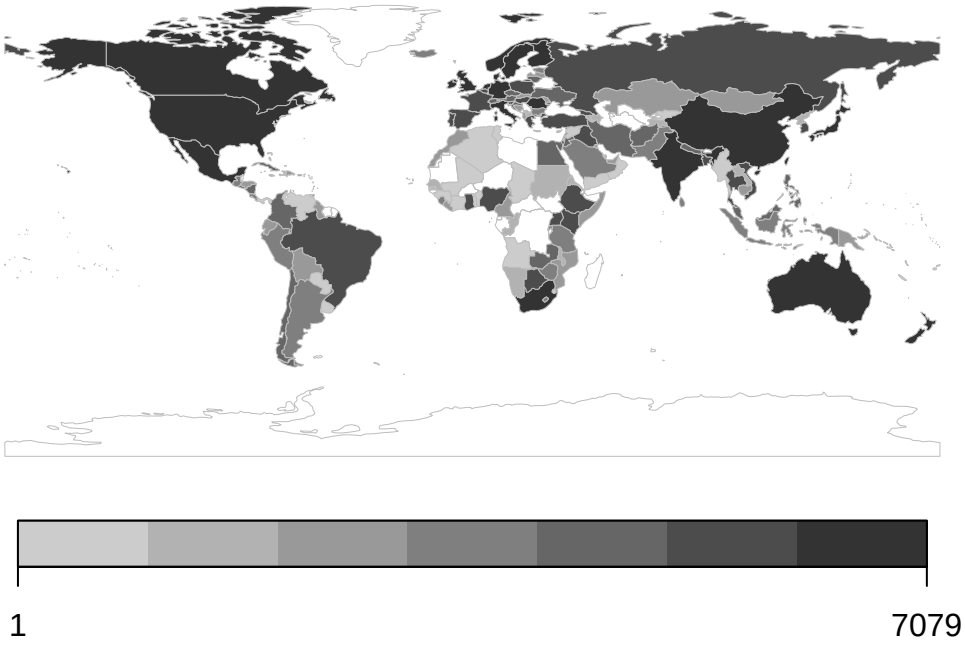
Practice: Social Work in Action	2005	2013	175	Y
Social Work in Public Health	2007	2013	157	Y
The Hong Kong Journal of Social Work	1993	2012	153	Y
Journal of Social Work in End-of-Life & Palliative Care	2005	2013	150	Y
School Social Work Journal	2002	2013	139	Y
Journal of Evidence-Based Social Work	2004	2013	123	Y
Journal of Social Work in Disability & Rehabilitation	2002	2013	115	Y
Ethics & Social Welfare	2009	2013	114	N
Journal of Social Work Research & Evaluation	2000	2005	98	Y
China Journal of Social Work	2009	2013	95	Y
Journal of Progressive Human Services	2000	2013	92	Y
Prevention in Human Services	1989	1993	90	N
Journal of the Society for Social Work & Research	2010	2013	63	N
Social Work Research & Abstracts	1989	1993	59	N
Canadian Social Work	1999	2005	47	Y
Journal of Applied Social Sciences	1994	1999	47	Y
Journal of Comparative Social Welfare	2008	2012	44	Y
Journal of Practice Teaching in Social Work & Health	1999	2005	39	Y
Rural Social Work	2000	2004	39	Y
Asia Pacific Journal of Social Work & Development	1997	2003	37	Y
Journal of Social Work Values & Ethics	2009	2012	33	Y
Journal of Practice Teaching in Social Work & Practice	1998	2000	27	N
Social Work Forum	2004	2009	27	Y
Contemporary Rural Social Work	2010	2013	23	N
Journal of Multicultural Social Work	1999	2000	21	N
Journal of Social Work & Human Sexuality	1989	1993	18	N


A. Number of Articles by Year
Articles
2000
1800
1600
1400
1200
1000
800
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
Year

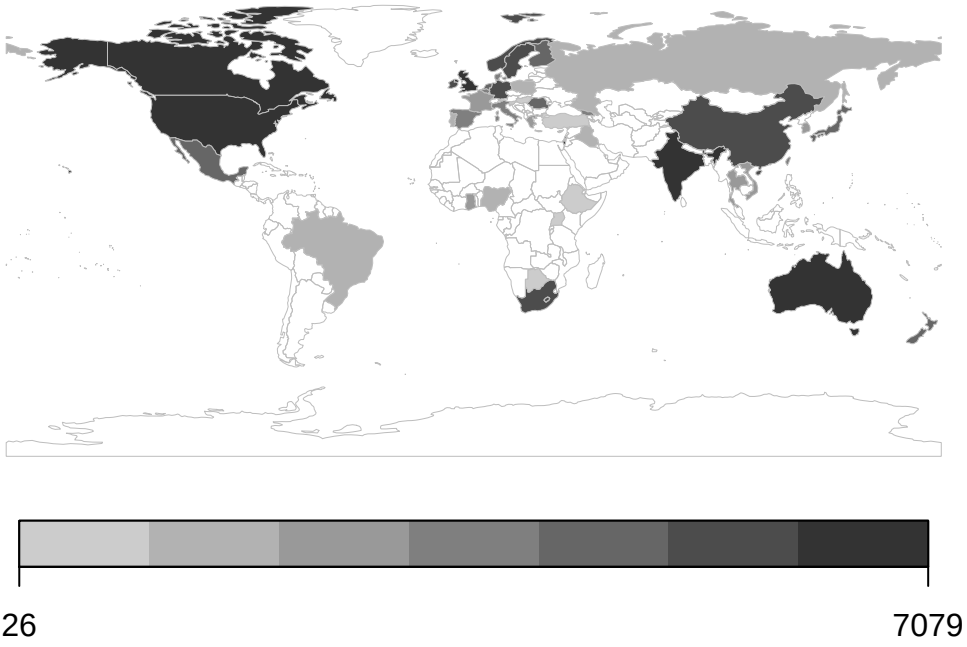

B. Number of Journals by Year
Journals
70
60
50
40
30
20
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
Year

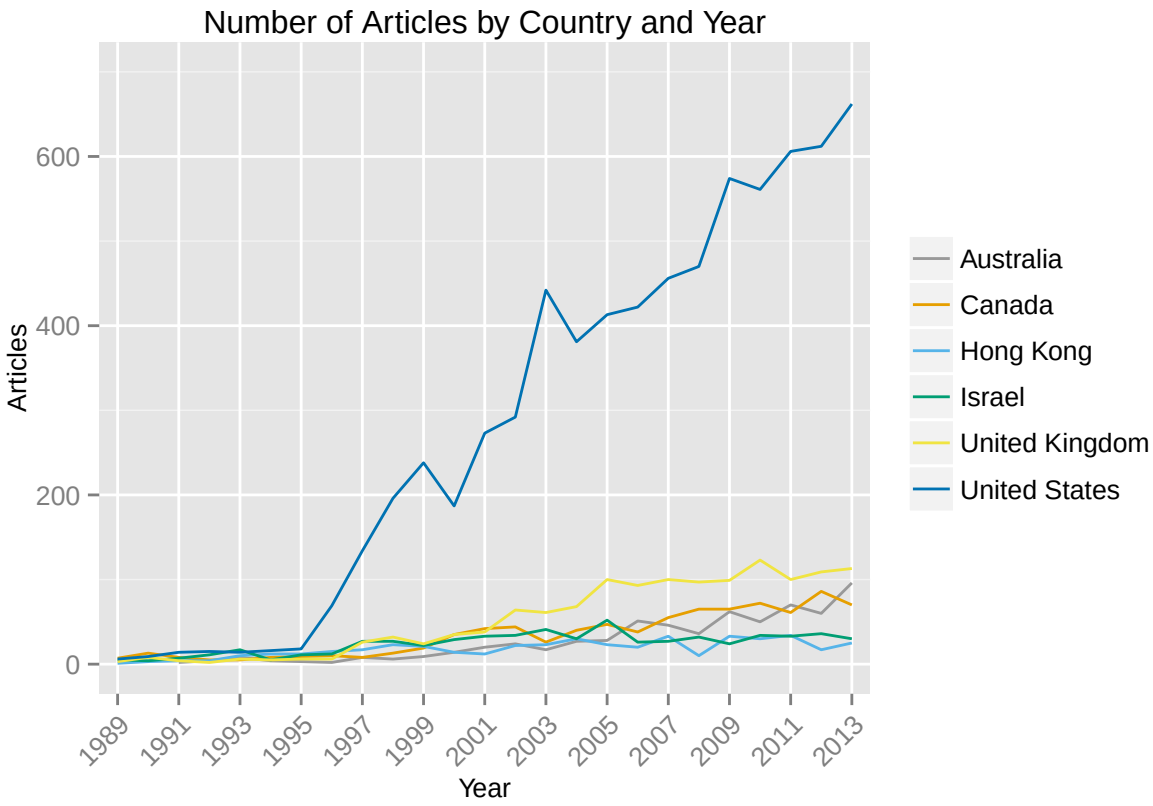

Number of Earned Social Work Doctoral Degrees by Year
Doctoral degrees
350
325
300
275
250
225
200
238
273
304
323
308
287
338
332
308
289
280
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
Year

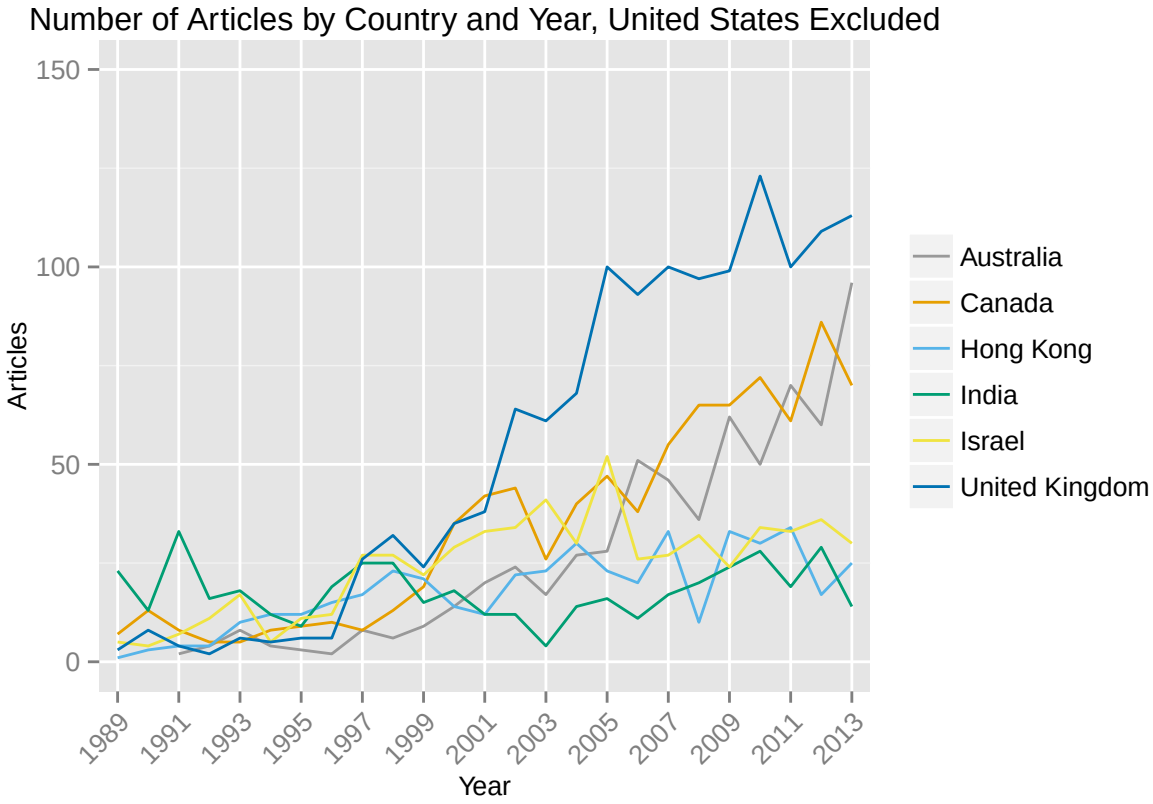
Article Counts by Country, 1989−2013

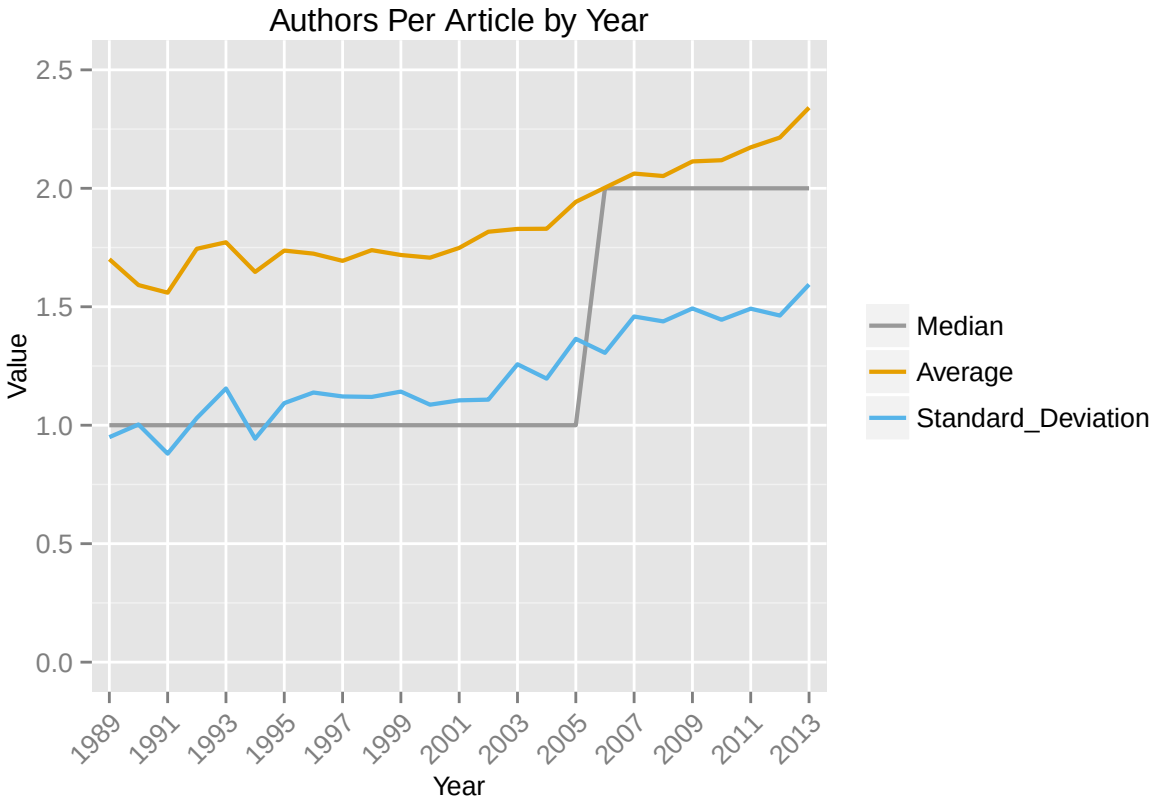


Article Counts for Countries with Sustained Presence, 1989−2013




Number of Articles by Country and Year
Articles
600
400
200
0
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
Year
Australia
Canada
Hong Kong
Israel
United Kingdom
United States


Number of Articles by Country and Year, United States Excluded
Articles
150
100
50
0
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
Year
Australia
Canada
Hong Kong
India
Israel
United Kingdom


Authors Per Article by Year
2.5
2.0
1.5
1.0
0.5
0.0
Value
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
Year
Median
Average
Standard_Deviation

Appendix A

Articles on Hodge and Lacasse (2011) but not in the current study.

```
 [1] "Annual of Social Work"
 [2] "Black Caucus"
 [3] "Caribbean Journal of Social Work"
 [4] "Critical Social Work"
 [5] "Electronic Journal of Social Work"
 [6] "IUC Journal of Social Work Theory & Practice"
 [7] "Japanese Journal of Social Services"
 [8] "Journal of Changsha Social Work"
 [9] "Journal of Forensic Social Work"
[10] "Journal of Rural Social Work & Social Development"
[11] "Journal of Social Work in Long-Term Care"
[12] "The New Social Worker"
[13] "The Social Worker/Le Travaileur Social"
[14] "The Spirituality & Social Work Forum"
```