A Follow-up Study on Data Management and Data Sharing Training in Graduate Education in the Social Sciences

Ashley Doonan and Evan Cosby
The Case for Graduate Training

• Need for researchers to be transparent
  • Sharing data is the best way to accomplish this (going beyond writing papers)

• Good data management is essential for good science
  • Replicability, secondary analysis, reducing cost and effort, etc.

• On the job training is insufficient – learning these skills needs to happen in graduate research training

• The role of data science professionals
Previously...

2016-2017 Survey of Graduate Programs

• Programs were confident in research skills (67%), 42% believed graduates have good data management skills

• 20% had an ethics course; only one had a data management course

• Nearly all programs required research methods course
  • Most programs reported data management included in methods
  • Less than half covered data sharing within the methods course

• Programs interested in repository programming
  • 62.5% maybe interested - 20.8% definite interest

(Ebersole & Holt, 2017)
Project Progression

1. Initial Program Survey
2. Syllabus Analysis
3. Expanded Program Survey
4. Graduate Student Survey
5. Cross Discipline Comparisons
Research Questions

• Objectively, what content are instructors covering in methodology courses?
  • What information do instructors present to graduate students when they are learning how to conduct research?

• Is material regarding data management or data sharing included in the overview of coursework?
  • Is it significant enough in the course to include?
  • Would it be obvious to someone outside of the program that students received instruction on this material?
Methods

1. Creating the Sampling Frame
2. Gathering Syllabi
3. Conducting Text Analysis
Methods - Sampling

• Seven fields included
  • Previous inclusion criteria: Social Science, research oriented, major governing body with an ethical code
  • Matching the previous six inclusions, with the addition of Economics

• A randomly selected list of U.S. programs was created using gradschools.com
  • Only programs with in-person courses were included
  • Mixed inclusion of both Master’s and Doctorate programs
Sample

- 140 graduate programs included in syllabus search
  - Random selection included 20 programs per field
  - Private and public universities
  - Schools of all sizes
  - All Carnegie classification levels included, as well as unclassified
  - Within all 50 states
Methods - Collection

- Required coursework was identified for programs using public webpages

- Syllabi were sought for required courses in the following areas:
  - Methodology, discipline specific statistics, discipline ethics, ethics in research, or similar

- Syllabi were found in databases such as OER Commons, or on the programs’ websites.

- Syllabi were available for 50 programs in total.
Methods – Available Syllabi

Bar chart showing the number of syllabi collected per field:

- Anthropology: 6
- Geography: 6
- History: 4
- Political Science: 10
- Psychology: 5
- Sociology: 9
- Economics: 10
Syllabi Source Characteristics

CATEGORIZED COURSE TYPE

- Research Methods/Methodology: 36%
- Discipline Statistics/Analysis: 34%
- Quantitative: 16%
- Qualitative: 4%
- Research Design: 10%
Syllabi Source Characteristics

• Course Requirements:
  • Completion of a research project (66%)
  • Collection of data (30%)

• Program Type
  • Masters (28%)
  • Doctoral (72%)
Methods – Text Analysis

- Search function and visual scan
  - Highlighting applicable words or phrases
  - Full document search
  - Headers and titles were excluded

- Allowing for permutations or implied mentions:
  - Synonyms or similar meanings were used
  - “Survey data collection, interpretation, and analysis” was 3 total mentions
Results

Mean # of Mentions

<table>
<thead>
<tr>
<th>Category</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Total</td>
<td>11.56</td>
<td>9.06</td>
</tr>
<tr>
<td>Collection</td>
<td>1.84</td>
<td>6.84</td>
</tr>
<tr>
<td>Analysis</td>
<td>6.56</td>
<td>6.84</td>
</tr>
<tr>
<td>Interpretation</td>
<td>3.84</td>
<td>3.58</td>
</tr>
<tr>
<td>Management</td>
<td>1.38</td>
<td>2.47</td>
</tr>
<tr>
<td>Sharing</td>
<td>.30</td>
<td>.84</td>
</tr>
</tbody>
</table>
Results

PROPORTIONAL MEAN INCLUSION OF DATA BY FIELD

- Sociology
- Psychology
- Political Science
- History
- Geography
- Economics
- Anthropology

Data Analysis  Data Presentation  Data Collection  Data Management  Data Sharing
Results

• Indications of Using or Interacting with an Archive, Database, or Library

<table>
<thead>
<tr>
<th>Total Mentions</th>
<th>Mean</th>
<th>SD</th>
<th>% Not Mentioned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archive Mentions</td>
<td>22</td>
<td>.44</td>
<td>1.64</td>
</tr>
<tr>
<td>Database Mentions</td>
<td>43</td>
<td>.92</td>
<td>1.99</td>
</tr>
<tr>
<td>Secondary Analysis</td>
<td>61</td>
<td>1.22</td>
<td>2.41</td>
</tr>
<tr>
<td>Data Archival</td>
<td>34</td>
<td>.68</td>
<td>1.72</td>
</tr>
</tbody>
</table>

• Any reference to an archive, library, or database (for any reason) in 22% of syllabi
Results

Inclusion of Ethics versus Statistics Content by Field

<table>
<thead>
<tr>
<th>Mean Number of Mentions per Syllabus</th>
<th>Ethics</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anthropology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Economics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geography</td>
<td></td>
<td></td>
</tr>
<tr>
<td>History</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sociology</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

18
Conclusions

• Previous survey suggested data management content would be represented in methods courses
  • Data management references appeared in limited amounts

• Training in data sharing or data repository use appears to be limited
  • Minimal references to archives or databases, even as resources

• Syllabi discuss statistics and performing analyses heavily, but how to care for data and comply with ethical requirements is often overlooked
Conclusions

• Is material regarding data management or data sharing included in the overview of coursework?
  Rarely, if at all
  • Is it significant enough in the course to include?
    Even when high level of collection detail is included, what to do with data after is not mentioned
  • Would it be obvious to someone outside of the program that students received instruction on this material?
    In the majority of cases, no
Possible Actions

• Supplemental coursework
  • Self-directed online content
  • Webinars or video series
  • Workshops
  • Use of example datasets with syntax
  • “Self-curation” in multiple stats packages

• Certification in data management skills
Discussion

• Providing language for professors to use in syllabi?
  • How to practice good data management or where to seek out help
  • Ethical importance of data sharing, how to share data, and how to find data

• Room for consultations or data management user support services?
Future Questions and Project Steps

1. Initial Program Survey
2. Syllabus Analysis
3. Expanded Program Survey
4. Graduate Student Survey
5. Cross Discipline Comparisons