Understanding selection bias, time-lags and measurement bias in secondary data sources: Putting the Encyclopedia of Associations database in broader context

Shaun Bevan a,*, Frank R. Baumgartner b, Erik W. Johnson c, John D. McCarthy d

a Mannheim Centre for European Social Research (MZES), University of Mannheim, Germany
b Department of Political Science, University of North Carolina, Chapel Hill, United States
c Department of Sociology, Washington State University, United States
d Department of Sociology and Criminology, Pennsylvania State University, United States

A R T I C L E  I N F O

Article history:
Received 1 January 2013
Revised 9 April 2013
Accepted 12 August 2013
Available online 22 August 2013

Keywords:
Secondary data
Measurement
Associations
Interest groups
Social movements

A B S T R A C T

Secondary data gathered for purposes other than research play an important role in the social sciences. A recent data release has made an important source of publicly available data on associational interests, the Encyclopedia of Associations (EA), readily accessible to scholars (www.policyagendas.org). In this paper we introduce these new data and systematically investigate issues of lag between events and subsequent reporting in the EA, as these have important but under-appreciated effects on time-series statistical models. We further analyze the accuracy and coverage of the database in numerous ways. Our study serves as a guide to potential users of this database, but we also reflect upon a number of issues that should concern all researchers who use secondary data such as newspaper records, IRS reports and FBI Uniform Crime Reports.

1. Introduction

Our aims in this paper are twofold. The first is to describe a systematic data set we have assembled that will allow researchers to track, over an extended period, the national associations in the US that make up what political scientists call the interest-group system or the associational universe. The source of that data is the Encyclopedia of Associations, a publication that was not originally assembled for research, but which has extensive research value nonetheless. The second is to situate the potential biases that we enumerate and assess in the EA into a broader context of problems of bias that have been investigated in the use of other widely used data sets, including FBI Uniform Crime Reports, IRS 990 filings, and newspaper reports of protest. We call such data secondary, in that is not initially collected by researchers for their own purposes, but distinguish it from the typical use of the term to describe survey data that is used by researchers who did not collect the survey data themselves. We begin by introducing the EA data set, then describe our assessment of its biases, and conclude with a discussion of the lessons we believe our efforts teach, comparing and contrasting our assessments of the EA with similar efforts to assess biases in newspaper, IRS and FBI crime data.

* Corresponding author. Address: University of Mannheim, MZES, D-68131 Mannheim, Germany. Fax: +49 621 181 2845.
E-mail address: shaun.bevan@gmail.com (S. Bevan).
2. The encyclopedia of associations

Since 1956 Gale Research (which became part of Thomson Learning in 1984, and a part of Cengage Learning in 2007) has published an annual volume listing US voluntary associations active at the national level: the Encyclopedia of Associations (EA). The EA is intended to act much like a trade directory that potential members and other interested parties can use to locate information about groups of interest. This directory is the publisher’s best attempt in any given year to capture the total number of national-level voluntary associations operating in the United States. The annual nature and consistently meticulous search process used by the compilers of the EA make it an excellent resource for the creation of a time-series dataset on national-level voluntary associations.

For each organization listed, the EA provides contact information, an abstract describing its activities, information about its membership size, staff, budget, and details such as whether it publishes magazines, holds annual conventions, etc. The EA is organized into 16 sections including Trade and Business, Health, Agriculture, Environment, Public Affairs, Social Welfare, among others, and each group is associated with a keyword that further highlights the main focus of its activities. Hundreds of highly detailed keywords are used. The 16 sections used to organize the Sections of the EA have been used in a consistent manner since 1959, with only minor changes to the subject names over time.

The EA’s empirical value has been demonstrated previously by other research, generally within specific areas of interest, e.g., the women’s and civil rights movements (Minkoff, 1995, 1997), gay/lesbian movement (Nownes, 2004) labor unions (Martin et al., 2006), human rights (Walker et al., 2010) and the environment (Johnson et al., 2010; Johnson and Frickel, 2011). Many scholars are interested in tracking populations of associations within policy sectors or issue-domains (Aldrich and Ruef, 2006; Davis et al., 2005; Hannan and Freeman, 1989; McCarthy and Zald, 1977; Minkoff and McCarthy, 2009; Perrow, 1992; Walker, 1983, 1991), and they use a variety of methods to do so. Important sources of data include general organizational directories like the EA and Yearbook of International Associations (see e.g. Boli and Thomas, 1997; Smith and Wiest, 2012), as well as domain specific directories (e.g. Edwards, 1994), IRS 990 forms (Brulle et al., 2007), activist biographies (Rosenthal et al., 1985), newspaper reports (Soule and King, 2008), and combinations of secondary sources and qualitative methods to build systematic enumerations of organizations within a defined geographical area (e.g. Andrews and Edwards, 2005; Kempton et al., 2001).

Despite the variety of strategies employed to identify relevant populations of social change organizations, limitations of data access have largely restricted previous analyses to single topics or to a single population or community of organizations. Building comprehensive knowledge from case studies one issue-domain at a time is a limited route to generalizable theory. Whereas information on commercial establishments, trade unions, and registered political parties, for example, is relatively plentiful and readily accessible, the lack of available data for systematic research on the characteristics and dynamics of broad associational populations has been the subject of repeated lament in social movements and interest group literatures (Gray and Lowery, 2000; Knoke, 1986; Lowery, 2012; McCarthy and Castelli, 2002; Schlozman and Tierney, 1986; Walker 1991). By aggregating and making available the full data associated with the entire set of organizations listed in the EA (Gale Research, 1970–2005), we aim to spur more comprehensive assessments of the population dynamics of US voluntary associations.

In 2012 the Policy Agendas Project (PAP; see www.policyagendas.org) made available annual counts of associations present per year in each of the PAP topic categories. The PAP also makes available time-series databases relating to national government activities of all types from 1948 to present. By linking the EA database with those of the PAP, we allow scholars to study the linkages between growth and development of the voluntary sector with that of government. As government has become more active in education policy, have more education-related associations formed? Or, were associations leading indicators, preceding government activities in various fields? The PAP database also provides a cross-walk through its issue codes to a database composed of all reports of protest in the New York Time between 1960 and 1995, inclusive (http://www.stanford.edu/group/collectiveaction/cgi-bin/drupal/), allowing for the first time systematic analyses of public protest, the formation of voluntary associations, newspaper coverage, and government activity not just for a single issue-domain, but in all areas of public policy.

Our intent here is to assess the quality of the database and to bring front-and-center what we have learned about the apparent lag between the time when a group forms and the time when it is reported in these published volumes. The lessons we have learned throughout our efforts to create a high quality database from an existing published source, we believe, should have broad resonance for any scholars interested in using secondary data of any type.

2.1. Creating a database

We began the daunting project of compiling a time-series database covering 35 years and more than 20,000 records per year by looking at two sections of the database: Labor Unions, and Public Affairs. The Labor Union section had fewer than 300 records and focusing on it allowed us to understand the logistics and difficulties of the large task we were undertaking.

---

1 Before 1974, volumes were published less regularly. Today, the printed version still appears each year, but many university libraries offer an electronically searchable database that is continuously updated: Associations Unlimited. The EA can be thought of as an annual snap shot of this database.

2 The only major change in the 1970–2005 time period is the addition of a section on Fan Clubs in 1987.

3 A bibliography including all social science journal articles using the source in some way generated from an exhaustive search of the JSTOR database is available (see Baumgartner and McCarthy, 2009).
Moving to the Public Affairs section allowed us to cover many politically relevant organizations, but still to be focused on fewer than 2000 groups in each volume rather than the full complement of organizations included in the EA. By beginning with a focus on two sections of the larger database we were able to draw some important lessons, which we incorporated as we expanded the project to the entire contents of the EA.

2.2. Accuracy and coverage

Several studies have addressed the issue of selection bias in the types of organizations included in the EA (e.g. Brulle et al., 2007; Johnson and Frickel, 2011, appendix A; Walker et al., 2011). We review this literature here, and our modest contribution, before moving onto the poorly understood issue of the lagged data structure present in these data.

In the most authoritative test of potential selection bias in the EA, Martin et al. (2006) compared the EA coverage of labor unions with the universalistic catalog provided by the US Department of Labor, which requires all labor unions to register with them and makes this information available to the public. They found that the EA included 95% of all labor unions having more than 10,000 members. Coverage was lower among very small unions, just 50% in the case of unions with fewer than 5000 members. Budgetary and staff data showed a similar pattern: coverage was near universal among large groups and lower for groups with fewer resources.

Other sections of the EA do not benefit from the presence of an official registry of national groups, so we cannot carry out a similar analysis of inclusion in the EA as a percentage of a national universe of known groups. We have strong reasons to believe, however, that the high level of coverage for large groups as well as the more spotty coverage of small or ephemeral groups is a general rule (see also Brulle et al., 2007; Minkoff, 1999). In later sections, we note how long it takes before a new organization is listed for the first time in the EA. This information about lag structures further corroborates the labor union analysis suggesting that large and stable groups are very highly covered whereas smaller groups, those without staff, and those that were only recently formed, are more likely to be missed. Scholars concerned with different elements of the US voluntary sector may have different reactions to this bias in coverage. For those interested in the impact of associations on public policy, it is certainly better to have high coverage of large (and presumably more influential) groups rather than small groups with no staff. For those interested in a fuller analysis of population dynamics, in particular the early stages of organizational life, the bias against small, new, and poor groups may be more problematic. For the typical user of the PAP, who may be interested in groups with the potential to affect public policy, omitting small groups with few resources is certainly preferable to missing the large, rich ones. Very few of these appear to be omitted from the EA.

3. Understanding lag structures

In a perfect world all new associations would be instantly recognized by the publisher of the EA, changes to association descriptions would be made in real time, and associations that fail would be removed from the directory immediately. Even in this hypothetical perfect world new associations and changes to the descriptions for existing associations could occur an entire year before being listed in a new volume of the EA, creating a lag between the event and its report. Here we pay careful attention to the delays between when an event occurs (e.g., a new organization is formed; it changes its budget, areas of activity, or staff size; or it goes out of existence) and when this new event is recorded in the annual EA. Understanding these delays is fundamental for scholars interested in time-series models of policy dynamics.

Available evidence suggests that the rules and procedures applied by EA editors to data collection are mostly consistent over time, with the reasonable exception of technological driven changes in search and information gathering procedures. Organizations are identified for inclusion primarily through scouring mass-media, with organizations then contacted directly with requests to provide the information necessary to complete EA entries. Minkoff found during her research in the 1980s that editors relied heavily on mass media (especially newspaper) accounts to initially identify organizations, and that once an organization was identified it typically took one additional year to contact the organization, assemble and publish relevant data (D. Minkoff, personal communication, March 28, 2013). This reliance on mass media coverage to initially identify organizations for inclusion in the EA existed both in earlier and, based on our conversations with EA staff, later time periods as well. The spread of digital media and new search technologies in recent years may be expected to reduce search costs and thus organizational listing lags (Minkoff, 2002). From our own statistical analyses we find that, on average, the printed volumes reflect events that occurred four years in the past for the time period covered by our data.

3.1. The lag between actions and their reporting

One bit of information recorded about each group listed in the EA is the date it was created. We can compare this reported creation date with the year when the organization first appears in the annual publication to assess lag structures. If a group reports being created in 1972 but does not appear in the EA before the 1980 edition, then we can say the lag is 8 years. About

---

4 Brulle et al. (2007) further suggest that groups located in or near Washington D.C. are over-represented in the EA data, though Martin et al. (2006) did not find this for labor unions. Bias in coverage favoring groups proximate to national policy makers enhances our confidence that policy-relevant national groups are largely listed within its pages.
90% of the groups listed in the *EA* include information on their creation-year. Fig. 1 shows the delay between creation and first listing, and the table below provides summary statistics.

Fig. 1 shows the difference between founding dates and first inclusion is typically about 4 years but there is a fat right tail to the distribution, with one group being listed in the directory 129 years after it was created. Our assessment is that these extreme outliers result from two processes. First, they result from the fact that many national associations were founded well before the *EA* began compiling its listing. Second, we suspect they stem also from changes in the purpose of the groups rather than deficiencies in the data collection process. This is the case with the group joining the directory 129 years after its founding, which was founded in 1865, but which originated as a state and later regional association. Recall that the *EA* is a directory of national-level voluntary associations. Many groups start out as local or regional groups and then, through merger, acquisition, or simple organizational transformation, become active on the national level years later. When asked to report the year they were created, they may well list the earliest relevant year, which might be when the largest parent organization was formed or during a period when the group was active only in a single state or region, and was therefore not eligible for inclusion in the *EA*. We do not always know exactly when a group shifted from being regional to having a national character, but our assessment of those groups with long delays in Fig. 1 suggests that most correspond to this process. Shifts in organizational scope are common as groups grow. This is why we present two sets of summary data: one for all groups listed and one excluding those groups with more than 10 years’ delay. The truncated distribution corresponds better to the delay from creation of a group to its first listing in the *EA*, we believe.

With these possible reasons for a lag between a group’s reported founding date and its year of first inclusion in mind, Fig. 1 can be understood more fully. What we can call the data collection lag is related to the process of surveying the landscape and identifying new organizations. The shifting scope of existing organizations (e.g. from a state-wide to national focus) may occur at any time in an organization’s life-cycle. The results in Fig. 1 are therefore a combination of the effects of the search process and the selection criteria. These two factors result in a Poisson distribution, making the use of the mean or median values a poor estimate of the expected difference between when a group is first included in the *EA* and when it first met the inclusion criteria. Even when only considering those groups with a ten year difference or less the mean and median values are inappropriate. It is safe to assume that as the number of years difference increases the percentage of differences due to the search process decreases, while the percentage of differences due to changes in organizational scope increases. The best single estimate of the difference between first inclusion and first eligibility is the mode: 4 years. Group counts based on the edition with a copyright year of 1990 therefore reflect group formations, deaths, and mergers from 1986, on average. As the figure makes clear, there is considerable variation around this average. Thus, we see an inherent measurement ambiguity.

---

5 Fig. 1 is based on the Public Affairs section of the *EA*. The number of groups per PAP topic code in the Public Affairs section correlates at more than 0.95 with the number of groups with that same topic code in the overall *EA* for the cases of civil rights, defense, government operations, and foreign affairs. Therefore we feel confident extrapolating from our more detailed analysis of the Public Affairs section.

6 A comparison of the most recent edition of the *EA* used in our analyses (copyright 2002) with the full dataset shows no significant differences over time in updating lag. The data for 2002 demonstrate the same second spike at 4 years as is displayed in Fig. 1. However, it is around this time that organizational web sites become commonplace (only a small percentage of our groups report websites in 2002, compared to near universal coverage in 2013). Future research should examine whether data-entry lags decrease in the most recent time period. More timely updating is exactly the intent of Associations Unlimited, the online and continuously update version of the hard-copy published *EA*. 

---

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>All groups</td>
<td>10.5</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>129</td>
</tr>
<tr>
<td>Excluding groups with more than 10 years of lag</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

**Fig. 1.** Count of differences between date first included and reported creation date.
relating to the original data collection process: some events are recorded in the database with a six-year lag, some after two years, and so on, with four years the typical case.

3.2. Updating an annual volume

We can assess lag structures in additional ways: by looking at how such information as budgets, staff sizes, and membership is reported from year to year for the same group, and by looking at what happens to listings after groups go defunct. It is clear from this analysis that while the compilers of the EA do their best to contact each group every year, in the absence of new information they continue to publish the most recent information previously reported. Thus, some percentage of what is listed as information for year \( t \) is actually about year \( t + x \), where \( x \) is a range of years, not a single value.

We provide here some examples of organizational updates that illustrate our concern. The American Political Science Association, for example, lists a membership of 2300 in the 1981 edition, 2400 in 1983 through 1987, then 2200 members in each volume from 1990 through 2003. The American Civil Liberties Union lists 275,000 members, 125 staff, and a budget of $20 million in each volume from 1998 through 2003. Our point is not that the directory is inaccurate, but that updates may or may not be made on an annual basis, depending on whether the group in question responded to the publisher’s inquiries or whether updated information was available for the publisher to update using public sources. Fig. 2 gives our assessment of how many years separate any changes in the numeric resources (that is, either budget, staff, or membership size) in the Public Affairs section of the EA. Note that this figure likely overestimates update lag, because we are unable to distinguish between the failure to update an organizational listing and instances where updates are made but there are no changes reported.

As Fig. 2 shows the compilers of the EA update resource information for most groups quite regularly; the modal category is every year. However, some groups clearly do not respond to the repeated requests they receive from the EA staff, and the distribution tails off to the right with some information remaining in print decades after it was last updated. The data in Fig. 2 indicate that updates follow a Weibull distribution and the mean and median values are therefore meaningful. The summary statistics reported in Fig. 2 indicate that group information is updated every 4 years on average with half of the groups being updated more frequently, but with some being updated much more rarely.

Fig. 2 is limited to those groups that continued in existence past the last year of observation. In Fig. 3 we present identical information for those groups that went out of existence. One might expect that for failing groups, one might see a longer period of inactivity before the compilers finally decided that the group no longer existed or received clear evidence of the group’s demise.

The results in Fig. 3 are quite similar to the results for continuing groups in Fig. 2 in the shape of the distribution, but there are some small differences. While the modal category for updates is again after a single year, a second peak in the data is found at the 4 year mark. Additionally, the proportion of updates occurring after more than 3 years is slightly higher with

---

7 However, as our update measures are a combination of staff, budget and membership numbers it is unlikely that these values would remain unchanged for five years in the majority of associations.

8 By failing we mean groups that were removed from the EA. In this sense a failure may mean several things the most common of which is the failure of the group outright, but also includes a group changing into a corporation or reverting back to a regional or even state association. However, all of these failures are by definition a failure for the group as a national-level voluntary association (see Bevan, 2013).
failing groups. Still, whether we look at continuing groups, failing groups, or the first appearance of newly created groups, Figs. 1–3 and the associated statistical summaries make clear that 4 years is the average lag from an event affecting a group and that information being reflected in the EA.

4. Imputing annual counts from five-year observations

Based on the analysis above, as we moved to create a comprehensive time-series of these large annual volumes, we determined that we could estimate the numbers of groups existing each year by PAP topic category reasonably accurately by assessing counts in the EA on the basis of five-year increments. We lost little in terms of measurement accuracy, since the entries are typically not updated annually in the first place. The PAP database relating to the EA therefore has an indicator variable coded 1 for those years where we compiled the data directly from the EA and coded 0 for those years which are estimates. Here we explain how we imputed the annual counts from the observations based on the five-year intervals.

Using an extended version of the PAP coding scheme (see Table 1) research assistants classified each association by topic. In order to create a consistent time series, we faced the task of linking the entry for a given organization, say the National Association of Manufacturers, in one year to the same organization in another year. We found that the purpose of an organization changed only very rarely, and such items as the creation date remained unchanged across all editions of the EA. So by matching each group to its entry in previous editions, we were able to create a single panel dataset. Creating this panel dataset was a time- and resource-intensive task, but the choice to code every five years was based on an investigation of the EA that revealed that less than 10% of all associations contained in the EA survive for fewer than five years. A five year sampling procedure assures that the vast majority of associations (greater than 90%) are captured in the EA dataset. Further, among organizations with significant staff, budget, and presence, virtually all survive for at least five years. The five year sampling process therefore produces a workable dataset of national-level voluntary associations in the US.

Table 1 presents the extension of the traditional PAP topic coding system, breaking apart Topic 12 (law and crime as topic 12, but with family-oriented organizations in their own category 11); sub-setting from public affairs groups (topic 20) those that were purely single issue-ideological groups (topic 32), and sub-setting from the general “culture” topic 23 those groups based on arts (topic 40), culture (41), and shared hobbies (42). The table shows the number of organizations in each of these extended topic areas for the 2005 edition of the EA. By consulting the PAP website users can easily see the number, per PAP topic category, for each year from 1970 to 2005.

While we believe the use of a five-year sampling procedure appropriately captures the population of groups contained in the EA it also makes necessary the imputation of data for the years not directly observed. We make use of our precise knowledge of the overall number of groups in each volume to improve on a simple linear interpolation. We count the precise

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>All groups</td>
<td>5.4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>Excluding groups with more than 10 years of lag</td>
<td>3.9</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

Fig. 3. Group updating time, failed groups.

9 For the labor union and public affairs sections, with which we started, we gathered information for each year, and the analysis above is based on the public affairs section of the dataset. After seeing these patterns, and the high cost of data collection in terms of person-hours, we determined it was better to complete the project based on observations at 5-year intervals.


11 See Honaker and King (2010) for a more general discussion of how to impute time series cross-sectional data. In this case, given the clear effect that the overall number of groups has on the EA dataset, we are confident that the method we discuss here performs well.
number in each annual volume, by EA section, and make that information available on the PAP web site. However, the EA sections do not correspond perfectly to the PAP topic codes, so our extrapolations are designed to calculate these estimates. The linear count of groups by issue is calculated for the missing data points by averaging the change in counts between each coded year. For example, if there are 100 groups in year and 120 groups in year+5 the growth of 20 groups is average over the four missing years producing imputed values of 104, 108, 112, 116 in order. We then transpose these annual counts of groups into percentages of each annual total. Since we know the actual total number, we adjust the estimated number of groups per topic code by multiplying the estimated percentage of the annual total by the actual total for that year. The resulting measure combines the linear imputation of group percentages by issue with the known count of groups and provides greater variance for the group count series based on real data. The resulting series by issue is presented in Fig. 4 in a stacked count graph.

Fig. 4 shows the dramatic changes that have taken place in the associational universe in the US from 1970 to 2005. Trade associations have always been an important element of the total, but their share has declined steadily as a proportion of the universe of national associations. Health, education, and environmental groups have grown quickly, while labor groups have stagnated. We do not elaborate on these substantive trends here (see Jordan et al., 2012), but the key take away is that this new database allows an analysis of change across the full range of national associations in the US over time.

Several scholars, as we have observed, have studied the growth of groups in a single issue domain or in a small number of related domains (e.g. Hannan and Freeman, 1989; Chong, 1991; Hansen, 1991; Minkoff, 1995, 1997; Nownes, 2004; Johnson and Frickel, 2011; Soule and King, 2008). Yet, surprisingly no scholar has systematically studied the growth of the group system in the United States over time despite the large body of work that considers that general process (e.g. Olson, 1965; Walker, 1983; Hannan and Freeman, 1989; Perrow, 1992; Aldrich and Ruef, 2006). The creation of the EA dataset and the imputations calculated from it will, we hope, help to change that by allowing for far more comprehensive investigations of groups across policy issues (e.g. Bevan, 2013).

### Table 1

<table>
<thead>
<tr>
<th>EA topic</th>
<th>PAP topic</th>
<th>Abbreviation</th>
<th>Name</th>
<th>Number listed, 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Economy</td>
<td>Macroeconomics, taxes, and the economy</td>
<td>83</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Civil</td>
<td>Civil Rights, Minority Issues and Civil Liberties</td>
<td>848</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Health</td>
<td>Health</td>
<td>2822</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Agriculture</td>
<td>Agriculture</td>
<td>987</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>Labor</td>
<td>Labor, Employment, and Immigration</td>
<td>418</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>Education</td>
<td>Formal Education System</td>
<td>1214</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>Environment</td>
<td>Environment</td>
<td>566</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>Energy</td>
<td>Energy</td>
<td>184</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>Transport</td>
<td>Transportation</td>
<td>790</td>
</tr>
<tr>
<td>11&lt;sup&gt;a&lt;/sup&gt;</td>
<td>12</td>
<td>Family</td>
<td>Family Issues</td>
<td>401</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>Law</td>
<td>Law and Crime Issues</td>
<td>507</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>Social</td>
<td>Social Welfare</td>
<td>467</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>Housing</td>
<td>Community Development and Housing Issues</td>
<td>192</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>Commerce</td>
<td>Banking, Finance, and Domestic Commerce including Business and Corporate Issues</td>
<td>2883</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>Defense</td>
<td>Defense</td>
<td>656</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>Science</td>
<td>Telecommunications, Mass Media, Space, Science and Technology</td>
<td>1151</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>Trade</td>
<td>Foreign Trade</td>
<td>156</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>Foreign</td>
<td>International Affairs and Foreign Aid</td>
<td>1321</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>Gov't</td>
<td>Federal Government, Public Policy Generally</td>
<td>424</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>Lands</td>
<td>Public Lands and Territories, Indian Affairs, Forest Management, and Government Dams, Water, and Irrigation Projects</td>
<td>151</td>
</tr>
<tr>
<td>24&lt;sup&gt;a&lt;/sup&gt;</td>
<td>24</td>
<td>State</td>
<td>State and Local Government and Policy</td>
<td>98</td>
</tr>
<tr>
<td>26&lt;sup&gt;a&lt;/sup&gt;</td>
<td>26</td>
<td>Weather</td>
<td>Weather and Natural Disasters</td>
<td>16</td>
</tr>
<tr>
<td>27&lt;sup&gt;a&lt;/sup&gt;</td>
<td>27</td>
<td>Fire</td>
<td>Fires</td>
<td>10</td>
</tr>
<tr>
<td>29&lt;sup&gt;a&lt;/sup&gt;</td>
<td>29</td>
<td>Sport</td>
<td>Sports and Recreation</td>
<td>795</td>
</tr>
<tr>
<td>30&lt;sup&gt;a&lt;/sup&gt;</td>
<td>30</td>
<td>Death</td>
<td>Death Notices</td>
<td>22</td>
</tr>
<tr>
<td>31&lt;sup&gt;a&lt;/sup&gt;</td>
<td>31</td>
<td>Church</td>
<td>Churches and Religion</td>
<td>943</td>
</tr>
<tr>
<td>32&lt;sup&gt;a&lt;/sup&gt;</td>
<td>20</td>
<td>Ideological</td>
<td>Ideological, Social Cause, and Political Groups</td>
<td>611</td>
</tr>
<tr>
<td>40&lt;sup&gt;a&lt;/sup&gt;</td>
<td>23</td>
<td>Arts</td>
<td>Performing, Fine and Creative Arts</td>
<td>1193</td>
</tr>
<tr>
<td>41&lt;sup&gt;b&lt;/sup&gt;</td>
<td>23</td>
<td>Culture</td>
<td>Culture, Heritage and History</td>
<td>1480</td>
</tr>
<tr>
<td>42&lt;sup&gt;b&lt;/sup&gt;</td>
<td>23</td>
<td>Hobbies</td>
<td>Hobbies, Collectors, Amusements and Clubs</td>
<td>940</td>
</tr>
<tr>
<td>99&lt;sup&gt;b&lt;/sup&gt;</td>
<td>99</td>
<td>Other</td>
<td>Other and Miscellaneous</td>
<td>95</td>
</tr>
</tbody>
</table>

Note that the PAP website lists all data with the traditional PAP codes but that the extended codes are available by downloading the full data file. N’s listed refer to the most recent, 2005, edition.

<sup>a</sup> An addition from the New York Times Codebook.

<sup>b</sup> An addition from the EA Codebook.
5. Lessons from similar secondary data sources for assessing bias in the EA

In any area of research, scholars using secondary data collected for other purposes must be concerned with its quality and potential biases. In this section we organize discussion around the quality of the source data as well as two sources of bias that are the focus of social science survey research more generally (it can be useful to think of the EA and similar directories as the result of a yearly organizational survey administered by the publisher): sampling bias and measurement bias. We seek to place bias contained in the EA within a broader perspective by focusing especially on how it compares with other secondary sources with similar attributes, but about which scientists know more. Knowledge from similar research communities on other popular sources of secondary data (such as the FBI Uniform Crime Reports, IRS 501(c)3 registrations and media), has heretofore failed to inform our understanding of increasingly popular organizational directory resources (including not just the EA, but also, e.g. the Yearbook of International Associations and the Directory of British Associations). These data sources are increasingly important within sociology, political science, and other social science research on a broad range of nonprofit organizations (including politically oriented groups), yet there is surprisingly little attention given to sources of potential bias in these directories. Research on organizational survey methodologies has similarly remained unconnected to the analysis of these organizational directories.

The reflections we offer concerning the EA database are meant not only to elucidate these issues for potential users of this valuable new resource, but to allow us the opportunity for a bit broader discussion of how to assess the quality of any similar secondary database. There are no doubt many other research communities working with other data sources collected for purposes beyond basic research. What is striking to us, however, is how investigations of bias in the sources we discuss here have run parallel to one another with almost no reference to researchers in other substantive domains who have considered very similar issues of bias.

5.1. Source quality

5.1.1. What was the original purpose for which the data was collected?

The first and most important question in using secondary data is also the most obvious, but is often overlooked. The purpose of the data and audience it was intended for play a large role in the data generating process. This can lead to inclusions or exclusions of potentially important observations, and introduce systematic biases in a secondary data source. Identifying these biases is a necessary first step in accounting for or correcting them.

Research on newspaper reporting of protest events is particularly instructive here. Media traces, particularly newspaper reports, have been widely used by researchers to study protest events (Earl et al., 2004). News reporters and editors very selectively determine what news items are published, governed by what systematic observers of the media have dubbed “news values.” Media gatekeepers are most certainly not influenced by the concerns of researchers who use their products, and the dimensions of selectivity of events to be covered (selection bias) and the content of media descriptions of those events (measurement of description bias) have been the focus of research aimed at delineating the biases in news products (e.g. McCarthy et al., 1996; Oliver and Myers, 1999; Oliver and Maney, 2000) demonstrating empirically the central role of news values in creating these biases.

In the case of the EA, the goal of the creators of the original volumes has consistently been to create a comprehensive list of national-level voluntary associations in the United States. The EA serves as a reference text for libraries and schools, as well as a source for businesses such as printers and event organizers seeking to identify potential clients. This suggests a strong motive for comprehensiveness exists for its creators at least when it comes to large and established groups.

Fig. 4. Imputed group densities by issue, 1970–2005.
the EA also lists information on name changes and group mergers, allowing users to trace individual groups over time as they evolve and change, suggests a strong motive for accuracy in these details. One popular alternative source of information on national non-profit associations is the Internal Revenue Service’s 990 registration and reporting system, which has become increasingly comprehensive in recent decades and now requires all 501(c) nonprofit organizations to register and submit annual financial reports that also include additional information on those organizations’ mission and annual activities. These reports are now finding wide use by researchers interested in the rapidly growing nonprofit sector in the US (e.g. Saxton and Benson, 2005). While this elaborate system was not established to gather data for researchers of nonprofit organizations, its evolution has been, in part, guided by such concerns through the efforts of the National Center for Charitable Statistics (NCCS, 2010a). Since the system is a voluntary reporting one, questions have been raised about the accuracy of information provided by nonprofits to the IRS (measurement bias) as well as the timeliness of registration (lag in reporting, a subcategory of selection bias), each of which will be discussed below.

Beginning in the 1920s the Federal Bureau of Investigation started collecting information about “crimes known” to the police. The mature FBI Uniform Crime Report (UCR) is the aggregation of reports from every police jurisdiction in the US on at least the seven major individual crimes (homicide, aggravated assault, forcible rape, burglary, robbery, theft and auto theft). The reporting system has achieved high rates of compliance, typically receiving reports from almost all jurisdictions annually. These official crime reports are often used to evaluate police performance among neighborhoods at the local level, and when made public provide opportunities to assess police performance between cities. These realities place heavy pressure upon police district commanders and agencies to underreport crimes (Mosher et al., 2002).

5.1.2. What is the reputation of the data source creator(s)?

The credibility of the data source is an important proxy indicator of quality. Established academic researchers and research institutions (e.g. NORC, Michigan’s Institute for Social Research/ICPSR, Gallup, etc.), for instance, have reputations to consider and can be expected to strive to produce the highest quality data possible that can, once archived, be used by other researchers. On the other hand, businesses (especially those dominated by short-term economic incentives) may produce data that is designed to suit their needs or frame the data in a way that serves those needs best. 12

In the case of the EA the publisher, Gale Research (founded in 1954, purchased by the Thomson Corporation in 1985, Cengage Learning in 2007), has throughout its history been focused on creating high quality reference sources and databases for sale to libraries and educational institutions. Founded by Frederick Gale Ruffner in 1954, the company quickly gained a stellar reputation for this task, building off the success of the EA. Gale Research is today responsible for producing a wide variety of reference texts beyond the EA, as well as respected online historical, newspaper and various other databases. Given its long-standing market niche as a producer of high-quality library reference works, the publisher can be expected to value the strength of its reputation and to strive to create as accurate and useful reference texts as possible. Furthermore, as a result of what we have determined is a common and consistent work culture at the publisher the same basic rules and principles have governed the gathering of data, except for reasonable changes in the search and information gathering procedures with the advent of new and emerging technologies. Personal meetings between a senior author of this article and the senior staff of Thompson / Gale responsible for compiling the EA made clear the high standards that the publisher sets as a continuing goal. Over the decades, compilers have developed considerable expertise and they appear to value their reputation for compiling high quality information. Were this reputation to be degraded, or should a rival publisher offer a more complete directory, considerable profit to the firm would be lost. 13

Among US researchers of protest events, the New York Times has been the primary source of data in large part because it is the premier national newspaper, with a reputation for objective reporting, an important rationale for its wide use as a research tool. The public availability of a comprehensive data set on protest events constructed from its reports 14 has quickened the pace of such research in recent years, and, as a result, we know most about biases in its coverage and reporting. The IRS has achieved a strong reputation for assembling information from tax records, although recent cuts in its agency funding are thought by some to have reduced its ability to audit reports, and, in general, its auditing attention to nonprofit reporting appears to be driven primarily by complaints from citizens and watchdog organizations. It does not yet possess a mechanism to enforce the recent registration requirement that all nonprofit organizations, even the smallest with minuscule budgets, register. The NCCS has a strong reputation as a vendor of statistical information on the nonprofit sector, including its management of the 990 data bases. The FBI devotes extensive resources to its UCR data gathering efforts, and is widely recognized to be doing a good job in monitoring local jurisdictions cooperation and encouraging compliance with its rigid reporting protocols.

5.2. Measurement bias

Measurement bias refers to the degree of accuracy of measures of the concepts they are intended to reflect. The publishers of the EA primarily rely on survey responses from representatives of individual groups to compile their volumes. The use of survey data from groups themselves means that the description of an association contained in the EA may be slanted toward representing the best public face of an association. One association in the dataset called the Salt Institute, for instance, states

---

12 See Lewis (2011) for stark evidence of this pattern by financial institutions leading up to and following the 2008 financial collapse.
13 The firm currently charges $1037 for the print version of Volume 1, and $2663 for Volumes 1–3.
14 http://www.stanford.edu/group/collectiveaction/cgi-bin/drupal/.
that its primary goal is to educate the public about the health benefits from a healthy sodium intake as well as the many common and uncommon uses for salt in our everyday lives. In reality the goal of the Salt Institute is not public education, but to promote the use of salt so that the many companies that are members of the group can expand their sales volume and thereby make greater profits. Research assistants in charge of coding the EA’s entries were aware of this incentive to misrepresent a group’s primary goals in the text of group abstracts and were trained to deal with them by using the implied meanings of the group abstracts rather than the literal meanings when appropriate. Thus we aimed to have groups like the Salt Institute be properly coded as commerce associations rather than health-focused ones. The EA dataset therefore represents an educated assessment of the primary public face of national-level voluntary associations over time.

Newspaper reports about the size of protest events are typically disputed by activist organizers who complain that they underestimate the number of demonstrators present, a reflection of the extent of support for the dissident message they intend to communicate (McPhail and McCarthy, 2004). While there are very few systematic assessments of such claims, one study suggests that newspaper reports are generally rather accurate in their reports of what journalists call “hard news,” that is the basic details about when and where an event occurred, its primary message, and the number of demonstrators present (McCarthy et al., 1998), lending credence to claims that measurement error in such reports is probably a trivial concern.

The wide use of IRS 990 filing information has prompted a spate of studies of the accuracy of financial information provided by the filings of tax-exempt nonprofits (Froelich et al., 2000; Yetman et al., 2009; Fischer et al., 2002). And, while discrepancies are revealed between voluntary reports and independent audits of samples of nonprofits, these appear, in general, to be relatively minor, leading the authors of one widely cited study to “conclude that the IRS 990 Return can be considered an adequate and reliable source of financial information for many types of investigations…” (Froelich et al., 2000, p. 232). We have been unable to locate studies assessing the accuracy of other forms of information included in the 990 returns.

5.3. Selection bias

Selection bias refers to biases in the inclusion of cases (individuals, organizations or events) in a data base. Newspaper selection bias has been widely studied by scholars of protest events and consistently demonstrates that large events are much more likely to be covered and smaller events much less likely to be covered. This finding mimics what we know about larger associations being more likely to be included in the EA, and about large nonprofit associations being almost universally registered and filing annual 990 reports while very small ones are far less likely to be registered and if so, less dutiful in filing annual reports.

Protest event selection bias studies have also investigated variation in patterns of bias between newspaper sources and mixed media sources, including comparisons of local with national sources, local with other local sources, and print with television sources. The empirical results of many studies have been mixed, sometimes showing major differences in source bias and sometimes relatively minor differences. 15 One strategy designed to generate more comprehensive event catalogs is the use of multiple news sources, a strategy that has become far more feasible with the wide availability of media source data in electronically searchable form (e.g. Schrodt, 2012).

Data bases typically are fraught with selection biases, many of which are intentional. As a result, alternative sources used for enumerations of organizational populations rarely include identical cases. That is, alternative sources can be expected to exhibit variable patterns of selection bias. The strategy of adopting multiple sources to triangulate a more comprehensive catalog of events in newspapers can be applied to the study of non-profit associations. It is likely that an enumeration strategy relying on a discrete number of broadly representative sources is the most feasible route towards a more comprehensive data-collection strategy. Clearly, the EA would be an important source for populations of national associations in the US.

Oftentimes, selection biases are the intentional result of explicit filtering rules. The intended coverage of the EA is every national-level voluntary association in the US active during the year of publication, for instance. Scholars interested in local groups, international groups, or non-membership lobbying organizations will find no useful information here. Knowing the intended coverage of a data source is a fundamental aspect of assessing its potential usefulness. Rules for including and excluding an observation from the data play an important role in the creation of a secondary data source and should be well noted by any researcher interested in using it.

Two main criteria must be met for an association to be included in the EA. The first is that an association is active at the national level, not just in certain regions or localities. The second criterion is that the association be a voluntary association, defined as an association whose membership joins by choice for some common goal or goals other than profit. Following these two criteria the publishers include a wide variety of groups from the Sierra Club to the United Autoworkers Union to Bobs International (an association comprised of people with the forename Robert that holds an annual national convention with awards that include the best shish ka-Bob). Despite including all groups matching the publisher’s definition of a national-level voluntary association by following these two criteria the publisher excludes the vast majority of formal organizations in the United States namely local/state groups and businesses. It is with these two criteria that the publisher of the EA has done a consistent job of documenting national-level voluntary associations over time.

15 One study in particular suggests that source selection bias makes newspaper protest event data highly unreliable for studying cross-community or time series patterns or protest (Ortiz et al., 2005).
A number of studies have compared different sources with regards to their coverage of associations and scholars have complained that different data sources lead to widely varying estimates of the number of groups. Scholars using various sources such as IRS Form 990s and enumerations of groups in a given locality (typically in a single issue domain such as the environment) sometimes identify numbers of groups that are very different from what would be found relying exclusively on the EA (Andrews and Edwards, 2005; Brulle et al., 2007; Carmichael et al., 2012). A particularly harsh critique of the EA as a comprehensive source of environmental advocacy organizations compares it with a variety of other listings of US national environmental organizations (Brulle et al., 2007; Carmichael et al., 2012), arguing that its lack of comprehensiveness renders it a poor resource for studying national environmental organizations. Two subsequent studies have addressed the question of selection bias in the EA, one for environmental organizations (Johnson and Frickel, 2011) and another for subcategories of advocacy groups (women’s, human rights, and civil rights) (Walker et al., 2010). Each makes a strong case for how to take known selection biases into account when interpreting models, and for the EA as a particularly valuable resource when analysts are interested in temporal change. While IRS data are more comprehensive in terms of the total volume of organizational listings, this source omits politically active groups that do not qualify for tax-exempt status. More troubling for analysts interested in temporal change is that, while IRS data are more comprehensive in terms of the total volume of organizational listings, this source omits politically active groups that do not qualify for tax-exempt status (Edwards, 1994; McCarthy et al., 1991). Biases contained within the IRS data do not include information on many other groups, however, since organizations that are not required to register with the government unless they seek formal non-profit status, and many (especially in the case of politically active groups) do not do so (Edwards, 1994; McCarthy et al., 1991). Biases contained within the EA privilege the types of organizations that are most central to the theoretical concerns of social movement and interest group scholars, those that are large and more politically active (as indicated by an inclusion bias towards groups located close to the nation’s capital). Relying on newspaper reports of groups would generate a much smaller list than what we have here, as the vast majority of the groups listed in the EA are utterly unworthy of extensive national news coverage (Amenta and Caren, 2010).

Our point is that differences in the number of groups observed by different methods of observation are not necessarily a sign of spotty or poor data collection practice; it could easily be because of different rules of coverage. Assessments of how well a given dataset measures its intended organizational universe must be made with reference to another observation of the same range universe, not by comparison with a database with a different intended coverage. Assessment of the labor

16 There is some evidence that non-response bias is less of a concern for surveys of organizations than it is for surveys of individuals (Smith, 1997).
17 See Maxfield (1999) for a description of the alternatives and a comparison between them.
union section of the EA (Martin et al., 2006) suggests that coverage for large, stable groups is virtually complete. It is less complete, however, for smaller and ephemeral groups. More important for some users, however, the intended coverage of the EA may not match what is needed.

5.3.1. Reporting lag: a sub-category of selection bias

Lags in reporting are another source of selection bias. The date of publication can serve as important signal not only about the content of the data source, but also quality. Many primary and secondary data sources are updated months or years after new activity. For example, budgets printed in the same year as they were issued are often estimates and revised numbers are published in the following year. This question on the date of publication often requires a broader look and analysis of the data than noting the copyright year in a published volume or the publication date of survey data. Surveys for instance take time to run and publishing a book of any sort also requires time. Field work dates and the point that data gathering stopped to finish a manuscript can both have large effects on the interpretation of data.

The question of when the data published or more importantly what was the observation point is especially important for time-series data. The data contained in the EA is in effect the best information available to the publisher at the time of publication. However, like any published volume there is a lag between event and publication due to the delays involved in data collection and publishing. The details of this lag and why it is a concern were discussed extensively in a previous section.

Similar issues of lags in the formal registration of a nonprofit organization with the IRS confront researchers using that data set for studying them. NCSS cautions users of its database, stating “The IRS Business Master Files include a field (RULEDATE) indicating when registered nonprofit organizations obtained formal recognition of their tax exempt status by the IRS (in other words, when the IRS approved their applications for exempt status). NCSS typically uses this as a proxy for when an organization was created. However, one should understand its origins and flaws before determining how best to use it in one’s research” (2010b). This problem is directly analogous to the lag structure we assess in the EA data set. It is not uncommon for there to be a significant lag between when organizations are created and when they apply for formal nonprofit status. We have not been successful in locating equivalent analyses for the IRS data base.

Protest event researchers also confront similar lags in the reporting of protest events. They typically frame narrow time windows around events seeking to locate reports, but often events are not reported for months and sometimes even years after they occurred. Widening the research time window can raise the costs of data collection considerably. The creators of the Dynamics of Collective Action data set were able to address this problem by examined all daily issues of the New York Times for a 35 year period, bundling all reports of an event no matter when they were made and centering them, on the date which they occurred. (See http://www.stanford.edu/group/collectiveaction/cgi-bin/drupal/ for a description of this element of their data collection methodology.)

6. Discussion

No primary or secondary data set is without its biases. An important goal of the user of any dataset is to determine the sources and the structure of its biases. The more we know about these biases the more we can adjust for them in our analyses and calibrate the amount of confidence to place in the patterns revealed by our analyses. This principle has motivated the extensive and careful work we have done to learn as much as we can about the EA. The EA database is a valuable tool for social science scholars interested in a range of substantive processes, and we trust future analysts will extend our inquiry into the structure and sources of EA bias and look into other secondary datasets as well.

Among the most serious biases we have identified in the EA database is its undercount of small associations and its lag structure. The EA systematically under represents small, understaffed, or ephemeral associations. For larger, significant associations the coverage is quite comprehensive. This has important implications for users; for some purposes the missing cases are relatively unimportant, but for other types of analysis this bias is prohibitive.

We focus on the less well understood, but significant, lag between new actions (such as the creation of an organization) and when these actions are reported in the dataset. This is our most important lesson: if the year of publication is 2005, the actions refer, on average, to 2001. But there is some inherent variance even here, so that some events listed as 2005 actually refer to 2002, 2000, etc., with 2001 being the single most common year.

The largest hurdle in using the EA dataset, therefore, is the establishment of the lag between the EA’s copyright year and the year it represents. Detailed analyses of the differences between group founding dates and first inclusion and of the average length of time it takes to update existing associations offer strong insights. Combined these analyses demonstrate that the data in the EA most closely matches the group system 4 years prior to the publication date. The nature of the EA makes it impossible to determine when every group was founded and failed. The 5 year sampling procedure employed by the project adds to this problem, but the lag structure of the EA itself means that problems with establishing the lag structure would exist whether the data was gathered every year, every 5 years or every 10. Calculating group counts by topic instead of focusing on the details of individual groups leads to more reliable aggregate values over time. This information will allow an analyst to take into account of lag structure in future analyses. Comparing the bias structures we have identified in the EA with those in three other widely used secondary data sets, we believe, puts them in the broader context of equivalent biases confronted by other research communities, and, in sum increasing our confidence in it.
While hesitant to make overly broad claims for data-sources not directly examined here, we believe our findings regarding the lag-structure of EA data has important implications for the large research literature relying not just on the EA but other similar organizational directories commonly employed as secondary data: including the *Yearbook of International Organizations* (e.g. Beckfield, 2010; Boli and Thomas, 1997; Frank et al., 2000; Smith and Wiest, 2012), *Directory of British Associations* (see Halpin and Jordan, 2012), and *Associations Canada* (e.g. Abelson and Carberry, 1998; Handy et al., 2008; Quarter et al., 2001). It is likely that all of these sources contain some sort of listing lag, although specifying the exact nature of that lag structure across these different data sources will require additional scholarly research.

The lagged structure of information contained within organizational directories, including that which we specify for the case of the EA, is more or less likely to be problematic depending on how information is accessed and used. A lagged data structure is relatively unproblematic for analyses which rely exclusively on organizational founding and mortality information contained in the EA to generate yearly organizational counts, as in classic organizational ecology models (e.g. Johnson and Frickel, 2011; Minkoff, 1997; Nownes, 2004; Walker et al., 2011). The lagged data structure is likely to be more problematic when the focus is on the precise timing of organizational change (e.g. Johnson, 2006; Minkoff, 1999), especially at the individual level. For aggregated counts of how many groups are active across various policy domains, as is the purpose of the PAP EA data release, the source is likely to be highly accurate, but they reflect a distributed lag of ± four years and this lag must clearly be understood by any user interested in time-series dynamics. While this has clear implications for time series models the same concerns hold for other quantitative and qualitative studies that also consider the relationship between a group or groups and other factors such as policy agendas, protests, public opinion, events and more. Without an accurate understanding of a group lifespans analyses run a very real risk of making inappropriate causal inferences.

More generally, our detailed analysis of the nature of errors of omission in the EA is not meant to discourage its use. On the contrary, we believe the database to be of great use to a wide range of political scientists, sociologists, and other academic audiences. But before scholars make use of secondary data resources, or any other, they should clearly understand the nature of any discrepancy between the social process “out there” in society, and its reflection in the archived records of the publisher of the dataset. While our theories tend to be about the social processes, our tests typically relate to the archived record. Some slippage between these two is typically inevitable, but all too often it is assumed to be limited, unbiased, and trivial. More often, it is an empirical matter that can be estimated directly and therefore understood.

**Acknowledgments**

The *Encyclopedia of Associations* data was gathered by Frank R. Baumgartner and John McCarthy through the National Science Foundation, “Collaborative Research: Database Development for the Study of Public Policy,” Grant Number SBR 0111611. Erik Johnson and Shaun Bevan served as lead Research Assistants on the project sequentially. The Policy Agendas Project data were originally collected by Frank R. Baumgartner and Bryan D. Jones, with the support of National Science Foundation Grant Numbers SBR 9320922 and 0111611, and are distributed through the Department of Government at the University of Texas at Austin. Neither the NSF nor the original collectors of the data bear any responsibility for the analysis reported here.

**References**


