1-2005

Using Archival Data for I-O Research: Advantages, Pitfalls, Sources, and Examples

Kenneth S. Shultz
California State University - San Bernardino

Calvin C. Hoffman
Alliant International University - Alhambra

Roni Reiter-Palmon
University of Nebraska at Omaha, rreiter-palmon@unomaha.edu

Using Archival Data for I-O Research: Advantages, Pitfalls, Sources, and Examples[1]

Kenneth S. Shultz
California State University, San Bernardino

Calvin C. Hoffman
Alliant University, Los Angeles

Roni Reiter-Palmon
University of Nebraska, Omaha

Two particular sets of experiences sparked our interest in writing this *TIP* article. The first was our increasing difficulty getting access to "new" organization-based samples. Depending on the topic and commitment involved, many organizations appear too leery and/or too strapped these days to allow for primary data collection. In addition, we have all experienced the disappointment of spending numerous hours on research proposals and meetings with organizational personnel, only to have the "plug pulled" at the last minute on a promising line of data collection. Conversely, we have also had experience with researchers in organizations who are willing and interested in partnering to analyze existing company data.

A second experience that sparked our interest was supervising graduate student theses and dissertations. Students likely have even more difficulty than faculty in gaining access to organization-based samples. As a result, they often end up collecting survey data on "working students" or other campus-based convenience samples. Although we realize that "working students" may often be appropriate subjects, depending on the research questions being asked, it has been our experience that students often resort to this strategy even when it may not be appropriate, once they find they can't obtain access to organization-based samples.

Given these experiences, we thought a short *TIP* article outlining some of the key issues of using archival data for I-O research would be of interest to many *TIP* readers. We by no means foresee (or propose) the use of archival data sets becoming the principal "data collection strategy" within I-O psychology. Rather, we see this as an underutilized tool to be added to current and future I-O psychologists' methodological toolbox. Given our extensive experiences working with a variety of sources of archival data, we realize there are numerous issues about which someone new to the area needs to be aware. Given the necessary brevity of a *TIP* article, we refer readers to key references cited throughout the rest of the paper for a detailed discussion of the issues raised below.

---

[1] This paper is based on a roundtable discussion at the Annual Conference of the Society for Industrial and Organizational Psychology (SIOP), San Diego, CA, April 2001. Address inquiries to kshultz@csusb.edu.

Brief Background on Using Archival Data

Researchers in many disciplines in the social sciences (including our closely related neighbors of economics and sociology) almost exclusively perform secondary analysis of existing data in their programs of research (Cherlin, 1991). Even within psychology, this issue is gaining more prominence. For example, in 1991 the journal *Developmental Psychology* had a special issue on secondary data analysis issues. Given that developmental psychologists are primarily interested in development changes over time, which ostensibly require longitudinal data, this should not be surprising (Brooks-Gunn, Phelps, & Elder, 1991; Duncan, 1991; McCall & Appelbaum, 1991).

Psychologists in general, however, appear to be reluctant to use existing data for research. Because our methodological training is almost exclusively geared toward the collection and analysis of new data, most psychologists do not consider using existing data to answer their research questions. If they do, they may be at a loss regarding where to start or what issues are of concern given their lack of training in using existing data.

Second, many psychology journals are leery of piecemeal publishing. As a result, many psychologists may view any reanalysis of existing data as simply piecemeal publishing. However, the APA publication manual (APA, 2001, p 353) clearly notes that:

> The prohibition of piecemeal publication does not preclude subsequent reanalysis of published data in light of new theories or methodologies if the reanalysis is clearly labeled as such. There may be times, especially in the instances of large-scale or multidisciplinary projects, when it is both necessary and appropriate to publish multiple reports…Repeated publication from a longitudinal study is often appropriate because the data from different times make unique scientific contributions.

Many large, nationally representative data sets are explicitly designed and collected with the intention they will be made available for public release and reanalysis by numerous scholars. Two examples include the Panel Study of Income Dynamics (PSID; Hill, 1992) and The National Opinion Research Center's (NORC) General Social Survey (GSS; Davis & Smith, 1992). The PSID is a longitudinal panel study that "gathers information about families and all individuals in those families through its annual interviews" (Hill, 1992, p. 7). Data collection was begun in 1967 on a nationally representative sample of 18,000 individuals. Data have been continuously collected almost every year since then on the same sample, which through marriage, divorces, remarriage, births, and so forth has now grown to over 40,000. Given the extensive nature of the data set, interested researchers can conduct cross-sectional, longitudinal, and/or intergenerational analyses using the PSID (Hill, 1992). PSID staff estimate that over 1,600 papers (including books, chapters, articles, working papers, government reports, and dissertations) have relied

on the PSID as the primary data for their research (C. Ward, personal communication, August 26, 1999).

The GSS, on the other hand, is an "almost annual" (see Davis & Smith, 1992) omnibus cross-sectional personal interview conducted by NORC. The first survey was done in 1972. In most years, a nationally representative sample of 1,500 individuals are surveyed, so over the years, more than 30,000 respondents have answered approximately 1,500 different questions. Davis and Smith (1992) report that approximately 2,000 books, articles, chapters, and dissertations have used the GSS as their primary source of data. Firebaugh (1997) presents numerous examples, including one on job satisfaction (e.g., Firebaugh & Harley, 1995), of how he used the GSS to analyze "social change" over time, which before his monograph, was generally thought to be impossible with repeated cross-sectional surveys.

Yet another source of data, and one which is also underused, is a strategy of using existing company databases, and comparing findings with those of published articles, and/or national or publisher databases. Cal Hoffman used this research strategy in a series of five articles published in *Personnel Psychology* (Hoffman, 1995; Hoffman, 1999; Hoffman, Holden, & Gale, 2000; Hoffman & McPhail, 1998; Hoffman & Thornton, 1997). For example, the 1997 paper contrasted assessment center and cognitive ability test data derived from two existing internal validation databases. The 1998 article contrasted existing company data from a PAQ job evaluation database against results published by Pearlman, Schmidt, and Hunter (1980). The 1999 paper dealt with physical ability testing, and again used an existing company PAQ database and the PAQ Services system database, coupled with results published by Blakely, Quinones, Crawford, and Jago (1994). In Hoffman et al., (2000) results from the company PAQ database and PAQ Services database were synthesized with results from nine internal company validation studies.

Methodological and Statistical Issues

Researchers using existing data sets must address numerous methodological and statistical issues. Although we clearly cannot address all issues in detail here, we will touch upon several and provide relevant references (e.g., Bryman, 1989; Elder, Pavalko, & Clipp, 1992; Finkel, 1995; Firebaugh, 1997; Kiecolt & Nathan, 1985; Lee, Forthofer, & Lorimor, 1989).

Clearly, reanalyzing existing data sets is not the only way of using existing data. Waldman and Avolio (1993), for example, discuss how researchers can use retrospective or postdictive research designs with archival data sets. We often hear calls for more longitudinal research in I-O psychology. One way to accomplish this would be to have researchers obtain archival data from organizations and supplement it by carrying out retrospective interviews or collecting follow-up primary data. This "new data" could be merged with archival data to create longitudinal data sets. I-O psychologists working in organiza-

tional settings have no doubt collected data for numerous cross-sectional studies that may have never been published because of a lack of "future data." Well the future is here, and many applied researchers and practitioners would likely be willing to help supplement such data sets with current primary data. Doing so would provide for much richer data sets than found in most cross-sectional studies and serve as excellent sources of data for theses and dissertations.

Use of existing data sets can also provide some significant methodological benefits. Using multiple existing data sets is an effective way to reduce, if not overcome, threats to internal validity like experimenter bias. Use of multiple data sets, or purely external data sets, is also a great way to bolster arguments about the generalizability of the results of a study. Finally, the convergence of findings from totally different databases collected by different researchers provides strong support for the construct validity of whatever it is you are reporting.

Potential Advantages and Pitfalls of Using Archival Data

The research process when using either existing or new (or some combination of) data is more similar than different, particularly at the beginning stages. No matter what the source of data, all sound research begins with an extensive review of the extant literature. Based on this review, hypotheses are formulated and reformulated. Once the research proposal stage is complete, the researcher may then begin to ask the question, "How best can I address my research questions and hypotheses?" In many instances, doing so requires collecting new data. In other instances, existing data may be available, either in its entirety or as a supplement to collecting new data, to adequately address such issues.

Table 1 outlines some of the key advantages and disadvantages of performing secondary analysis of existing data. The salience of the advantages and disadvantages depends on a variety of factors. For example, as a student, resources savings and easy access to existing organizational data may be key. As a professor at an undergraduate teaching institution where research assistants are few and far between, having data that is SPSS or SAS ready and being able to have instant access to longitudinal data may be a key factor. As an organizational-based researcher, being able to use existing company data as pilot data to justify a proposed organizational intervention may be the most salient factor for using existing data.

On the other hand, faculty may be leery of students using existing data for a fear of dustbowl empiricism or a stagnation of theory, and organizational-based researchers may not be as familiar with the unique statistical skills needed to complete such research and analyses. No matter what your position, you must weigh the various potential advantages and disadvantages outlined in Table 1 to determine if, for a particular situation, it makes sense to employ existing data. Either as the sole source of data or as a supplement or pilot to enhance future data collection, use of existing data must be well justified.

Table 1

Advantages and Disadvantages of Performing Secondary Analysis on Archival Data

Potential advantages	Potential disadvantages
❑ Resources savings	❑ Appropriateness of data
❑ Circumvent data collection woes	❑ Completeness of documentation
❑ A variety of research designs possible	❑ Detecting errors/sources often difficult if not impossible
❑ Usually SPSS or SAS ready	❑ Overall quality of data
❑ Relative ease of data transfer and storage	❑ Stagnation of theory
❑ Use as pilot data/exploratory study	❑ Lure of dustbowl empiricism
❑ Typically much larger and often national samples, as a result, can perform newer and more powerful statistics	❑ Unique statistical skills required
❑ Availability of longitudinal data	❑ Illusion of quick and easy research
❑ Availability of international/ cross-cultural data	❑ Convincing editors or thesis/dissertation advisors you are not simply duplicating existing research
❑ Organizations may be more open to using existing data versus collecting new data	❑ Failure of students to develop skills required in planning and conducting data collection

Assuming you decide that existing data may be a legitimate option, where does one get such data? Table 2 outlines some of the key sources of potential data. These sources include academic archives, government archives, private foundations, private and public sector organizations, and other independent researchers. Colleagues in related disciplines such as sociology and economics may be able to point you toward appropriate places for the former three sources, and fellow I-O colleagues would be the key resources to obtain data from the latter two sources.

Summary and Conclusion

In summary, we believe there has been an underutilization of archival data in I-O research. We believe it is the quality of the research questions, and the ability of the data to answer those questions, that should be of primary concern to I-O psychologists. Hence, I-O researchers may not need to collect new data to answer important research questions if existing data are available to do so. We must reiterate that we do not see the reanalysis of existing data becoming the dominant mode of "data collection" (as it is in other social science

disciplines such as economics and sociology). Rather, we wish to highlight its potential, while at the same time making it clear to those interested in using this strategy that it is not a panacea to avoid primary data collection and that it has many unique methodological concerns that must be attended to.

Table 2
Where to Obtain Archival Data

- Academic archives (e.g., ICPSR, DPLS, NORC – See Web links below)
- Government archives (e.g., Census Bureau, Department of Labor, military)
- Private/public organizations and consulting firms
- Private foundations (e.g., the Families and Work Institute—See Web link below)
- Other independent researchers

ICPSR: *Inter-university Consortium for Political and Social Research*—Most major universities in the United States and Canada (and throughout the world) have access to this extensive archive of over 20,000 data sets.

Started in 1962 at University of Michigan, Largest archive of computer readable data files in the world (~20,000 from 150 countries)

A few data archives to start with on the World Wide Web
ICPSR: http://www.icpsr.umich.edu/
Data and Program Library Service: http://dpls.dacc.wisc.edu/
Henry A. Murray Research Center: http://www.radcliffe.edu/murray
Families and Work Institute: http://www.familiesandwork.or

References[2]

APA (2001). *Publication manual of the American Psychological Association, 5th edition.* Washington, DC: APA.

Blakley, B. R., Quinones, M. A., Crawford, M. S., & Jago, I. A. (1994). The validity of isometric strength tests. *Personnel Psychology, 47*, 247–274.

Brooks-Gunn, J., Phelps, E., & Elder, G. H., Jr. (1991). Studying lives through time: Secondary data analysis in developmental psychology. *Developmental Psychology, 27*, 899–910.

Bryman, A. (1989). Archival research and secondary analysis of survey research (Ch #7, pp. 188–206). In *Research Methods and Organizational Studies*. London: Unwin Hyman.

Cherlin, A. (1991). On analyzing other people's data. *Developmental Psychology, 27*, 946–948.

Davis, J. A., & Smith, T. W. (1992). *The NORC General Social Survey: A user's guide*. Newbury Park, CA: Sage.

Duncan, G. J. (1991). Made in heaven: Secondary data analysis and interdisciplinary collaborations. *Developmental Psychology, 27*, 949–951.

[2] A listing of additional empirical papers that demonstrate the use of archival data in I-O research, which have been presented and/or published by the three authors and/or their students, is available from the first author at kshultz@csusb.edu.

Elder, Jr., G. H., Pavalko, E. K., & Clipp, E. C. (1992). *Working with archival data: Studying lives.* Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-088. Newbury Park, CA: Sage.

Finkel, S. E. (1995). *Causal analysis with panel data.* Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-105. Thousand Oaks, CA: Sage.

Firebaugh, G. (1997). *Analyzing repeated surveys.* Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-115. Thousand Oaks, CA: Sage.

Firebaugh, G., & Harley, B. (1995). Trends in job satisfaction in the United States by race, gender and type of occupation. In R.L. Simpson and I. H. Simpson (Eds.), *Research in the sociology of work, vol. 5: The meaning of work* (pp. 87–104). Greenwich, CT: JAI.

Hill, M. S. (1992). *The panel study of income dynamics: A user's guide*. Newbury Park, CA: Sage.

Hoffman, C. C. (1995). Applying range restriction corrections using published norms: Three case studies. *Personnel Psychology, 48*, 913–923.

Hoffman, C. C. (1999). Generalizing physical ability test validity: A case study using transportability, validity generalization, and construct validity evidence. *Personnel Psychology, 52,* 1019–1041.

Hoffman, C. C., Holden, L. M, & Gale, E. (2000). So many jobs, so little "*n*": Applying expanded validation models to support generalization of cognitive ability. *Personnel Psychology, 53*, 955–991.

Hoffman, C. C., & McPhail S. M. (1998). Exploring options for supporting test use in situations precluding local validation. *Personnel Psychology, 51*, 987–1003.

Hoffman, C. C., & Thornton, G. C. III. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology, 50*, 455–470.

Kiecolt, K. J., & Nathan, L. E. (1985). *Secondary analysis of survey data.* Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-053. Newbury Park, CA: Sage.

Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). *Analyzing complex survey data.* Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07–71. Thousand Oaks, CA: Sage.

McCall, R. B., & Appelbaum, M. I. (1991). Some issues of conducting secondary analysis. *Developmental Psychology, 27*, 911–917.

Pearlman, K., Schmidt, F. L, & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*, 373–406.

Waldman, D. A., & Avolio, B. J. (1993). Aging and work performance in perspective: Contextual and developmental considerations. *Research in Personnel and Human Resources Management, 11*, 133–162.