Educational Research and Reviews Vol. 6(20), pp. 993-996, 5 December, 2011
Available online at http://www.academicjournals.org/ERR
DOI: 10.5897/ERR10.130
ISSN 1990-3839 ©2011 Academic Journals

Review

Research using government data sets: An underutilised resource

Sally Knipe

Faculty of Education, Murray School of Education, P. O. Box 789, Albury NSW 2640, Australia. E-mail: sknipe@csu.edu.au Tel: 02 60519409.

Accepted 20 November, 2011

The use of existing data for education research activities can be a valuable resource. Improvement in statistical analysis and data management and retrieval techniques, as well as access to government data bases, has expanded opportunities for researchers seeking to investigate issues that are institutional in nature, such as participation patterns for young people transitioning from school to employment or training. Existing data may be used for descriptive research, correlation, cross-sectional or causal-comparative designs. Government institutions gather considerable amounts of data that could be 'mined' by researchers; however, the way that data are compiled and differences between data bases is a difficulty to address. This article argues for greater use of existing government data bases as a data source for research in education and the social sciences, and provides some insights for researchers to consider.

Key words: Ex-post facto, data mining.

INTRODUCTION

Most educational research is undertaken in natural settings, such as schools and classrooms, and researchers endeavour to gather data appropriate to the research questions being investigated concerning aspects of education. The source and quality of data gathered is a contributing factor to internal and external validity of research design. However, data that already exists can be a useful source, provided that the mining or collecting of the data are undertaken with a clear understanding of the possibilities and limitations of the information gathered and analysed. Indeed, for investigations that seek to describe and explain issues that involve institutions and government policy, information gathered by government departments, departments of education, and education institutions, would be an appropriate data resource. Researchers should make greater use of existing data stored in large government data bases as a data source for research in education and the social sciences.

Existing data as a data source

Accessing data in government data bases enables a researcher to extract data relating to a set of variables to describe the characteristics of a particular group of people. On the other hand, a researcher may analyse the data and explore relationships between variables in order to investigate possible pre-existing situations that may have caused subsequent differences between selected groups. Alternatively, a researcher may examine the data in order to undertake longitudinal cross-sectional research. Government data sets are a rich source of existing data for a variety research purposes.

Research that uses data that exists already, in order to investigate a research problem, is *ex-post facto* in nature (Wiersma and Jurs, 2005). The term '*ex-post facto* research' or '*ex-post facto* method' generally refers to causal-comparative research design, whereby an investigator searches through the data in order to gather information concerning the possible causes for some designated situation. Causal-comparative research provides information about the sequences and patterns within data related to phenomena (Isaac and Michael, 1997: 54). In causal-comparative research, using existing data, the researcher attempts to find out if differences between groups have resulted in an observed difference for an independent variable (Schumacher and McMillan,

1993: 285).

In the case of causal-comparative research method the researcher looks for cause and effect relationships but is unable to control variables as the data have been already collected. Where a cause and effect relationship between two variables is found, the researcher does not know which variable is the cause and which variable is the effect, merely that there is a relationship that could be investigated further. As well, the researcher can investigate similar relationships in order to build a set of plausible explanations of the research problem. Researchers from various disciplines, such as Rew et al. (2000) and Castle (2003) from the nursing sector, as well as Gorard (2001) in education, advocate the use of causal-comparative research methods using existing data, contending that existing data provides a viable form of analysis that can be used by researchers to respond to theories and test hypotheses.

Developmental research investigates change over time, and data may be gathered longitudinally to measure the rate of change in a set of subjects, or cross-sectionally to measure rate of change in a sample of representative time spans, or trend studies designed to identify patterns of change that occurred in the past (Isaacs and Michael, 1997: 50).

Developmental research may gather existing data from large government data sets. Oshima and Domaleski (2006) investigated differences in academic performance for children from kindergarten to grade eight in order to determine the effect of being older or younger in their kindergarten cohort. The researchers used existing data from a longitudinal study to identify a sample of children, together with data from state-wide criterion referenced testing, in order to gather information about academic achievement.

Where there is access to existing data that contains two or more selected groups and information related to variables of interest to a researcher, a causal-comparative method can be applied. The use of existing data in social and behavioural studies can be a way of relating patterns inherent in the data and making assumptions or deductions about aspects of society.

Existing data can be used for research designs that use a descriptive method or a causal-comparative method, or developmental or correlational research designs. Using the term '*ex post facto* research', rather than '*ex post facto* data' limits understanding of the possibilities for using existing data with a range of research designs; the emphasis should be upon the use of existing data rather than the research design (Wiersma and Jurs, 2005). The limitations of using existing data, is the fact that the researcher did not collect the data and the parameters are often not indicated.

Government data-sets as a data source

A range of organisations around the world collect data, such as the Organisation for Economic Co-operation and Development (OECD); the United States Census Bureau (USCD); and, the World Health Organisation (WHO). In Australia, significant amounts of data are collected by government and non-government organisations including the Australian Bureau of Statistics (ABS); the Australian Institute for Health and Welfare (AIHW), Australian Social Science Data Archive (ASSDA) and the National Centre for Vocational Education Research (NCVER). Within the education sector, organisations such as the Australian Bureau of Statistics collect data related to retention and attendance rates at school; and in the tertiary sector; the number of teachers employed in government and non-government schools; and adult literacy skills. To facilitate issues highlighted in this article, the Australian Bureau of Statistics data collection and administration will be used as an example.

The amount of time and cost saved in using existing government and non-government data could be of benefit to researchers if data available were relevant to an investigation. A limitation in using existing data concerns the way that the data were collected. However, assuming, for a particular data base, the design of procedures employed in gathering data was robust, the sample available to researchers would be much larger than samples usually available to researchers who gather data for a research project (Miller and Han, 2000). As well, at a cost, a researcher may choose to commission from a government organisation, such as the Australian Bureau of Statistics, provision of specific data that is relevant to a particular research problem but not readily accessible to the public.

Large data sets could be a useful data source for research that seeks to provide an overall picture of social patterns and trends to describe a particular situation. McMillan et al. (2009) and Jones and McMillan (2001) extracted existing data from the ABS and developed occupational scales drawn from the occupational classifications produced by the ABS, and then used these scales to determine occupational status scores for males and females in full-time and part-time employment.

Improvement in data management techniques and statistical data analysis software packages has served to make studies using existing data more justifiable. Further, data compiled by government organisations is now more accessible to a researcher through improved systems of computer storage and retrieval of data, as well as access to data through internet websites and computer software that enables a researcher to access, compress, and download data to a personal computer (Miller and Han, 2000).

During the 1990s, business, government, and scientific organisations, compiled larger and larger amounts of data in order to gather information about people that could be used for marketing purposes, policy development or research. Conventional methods of data analysis applied to large data bases were laborious and

time inefficient, which generated the development of ‘data-mining’ techniques more appropriate to analysing large amounts of data. Searching for patterns and trends in large amounts of data was considered to be a process of ‘knowledge discovery’, and ‘data-mining’ was described as the technique used for this process (Fayyad et al., 1996: Holscheimer, 1994). Data-mining is applied to a wide range of data analysis techniques used with large amounts of data.

In the past, research using large data bases was mainly undertaken by organisations such as the Australian Council for Educational Research, the National Centre for Social and Economic Modelling, and many research centres at universities. With improved and expanded access to government databases, such as ABS, research using government data sets is now within the realm of individual researchers in the social sciences and education. The development of websites such as the ‘home interaction program for parents and youngsters (HIPPY) Australia’, ‘Public Health Information Development Unit’, “My School” and the recently launched Australian Early Childhood Index (ACI), provide useful data sources for research activity in education and the social sciences.

Accessing government data bases

Government departments differ in the way they provide data for public access. Previously, summaries of data sets could be found as summary data in various publications, such as government reports and print material in the form of books and catalogues held in libraries that gave researchers an indication of the type and nature of data available. Organisation such as the ABS produce a wide range of materials, including catalogues, data sets, media releases, reports published and distributed in book form, and data in computerised aggregate format. In recent years, the ABS has placed most of its publications on the ABS Website and has made raw data available to the public through the ABS *Time Series Spreadsheets,* table builder and ‘data cubes’.

The ABS program *SuperTABLE* is used to extract raw data relevant to a particular variable or category, which can be down-loaded and converted into an Excel format. Super TABLES are a free web-based resource that enables users to view and manipulate data in multidimensional tables and allowing users to select individual data items. The process involves locating the required data file, downloading the file by using the (free) Space-Time Research software into SuperTABLE, ready for loading variables, for example, age, sex, state/territory, month/year. When data are compiled into tables and converted into an *Excel* spreadsheet, variables can be compiled into other spreadsheets (http://www.abs.gov.au/).

Data Cubes describe *Time Series* demographic variables, such as sex, age, and month, year, as a two or three dimensional collection of values. ‘Data cubes’ are used as a way of providing access to information stored in large data sets, and data held in these files can be also transferred into Excel spreadsheets. “... data cubes contain detailed tables for the data that underpins the tables, graphs and commentary included in the summary information sections” (http://www.abs.gov.au/). Organisations such as the national centre for vocational education research (NCVER) and the ABS use ‘data Cubes’ to store national data.

The ABS *TableBuilder* is an online tool that allows data to be selected and collated into a table from the ‘census output record file’ that is related to a range of geographic areas. Data is drawn from the Census, which is undertaken every five years in Australia, from a range of categories such as, education, housing, income, transport, religion, ethnicity, occupation, family composition location. Once the required data are complied data can be also transferred into excel spreadsheets http://www.abs.gov.au/tablebuilder

Multi-site comparison of data

Government agencies collect and report data in a number of ways using different population profiles, and access to large government data bases has become easier with advancements in storage and retrieval of national data. Researchers using one large data base access and compare data derived from within that data base, according to the definitions used in compiling data. However, researchers investigating a research problem that involved accessing different data sets would need to determine the comparability of the data that they were seeking to use. Multi-site data comparison may range from drawing data from different data sets that is comparable (for example gender) to drawing data that is not compatible and converting the data into a format whereby data comparison can be justifiable (Knipe, 2009).

Data contained in large data sets is likely to be reported using a variety of variables that may be presented in several formats. For example, the reporting age groups may be by an individual age or by an age grouping. Also, procedures for data collection may change as a result of policy or legislation so that data collected for one year may not be consistent with data that were collected for another year, eg definitions for unemployment or disability pension recipients. Multi-site comparison of data undertaken by a researcher can identify and clarify categories of data that are consistent and can be used for data analysis, and this process may reveal inconsistencies between categories that could be a guide for further research.

Probably, the ABS has been the most readily used

Australian data base available to the public over time and, despite some changes, has maintained a consistent set of categories for data collection. Variations in the way that data are compiled should not deter researchers from accessing a rich source of existing data available for research purposes. Anomalies identified in large data bases that are used to inform policy decision-making would be an important research finding.

Conclusion

Although the format of government data sets may present a challenge for a researcher, existing data available from data sets, drawn from government data bases collected by government departments, offers a rich source of information that could be used for investigation of educational issues. Using existing data in a research study, either for a causal-comparative study, a descriptive study, a developmental or a correlational study, is an underutilised approach in education research. With large data sets now more easily available as a source of data, and improved methods of statistical analysis to analyse such data, researchers should consider the usefulness of this resource. The use of data from government data sets is appropriate where a research problem has a national focus and/or straddles different government institutions, such as universities, schools, and TAFE colleges or non government education providers.

Researchers using one government data base would be able to access and compare data derived from within that data base. However, researchers investigating a research problem that was cross-institutional in nature, and drew data from different data bases, would need to determine the comparability or otherwise of the data that they were seeking to use, and to recognise limits in comparability of data relating to particular variables, in which case multi-site data verification could be applied. The vast amounts of information that are available in a government data base is a useful and useable resource for research purposes, and well within the capacity of individual researchers. Improvements in computer technology have made the mining of data easier and more accessible.

Where data sets are linked to an individual's personal detail, accessing this data for research purposes is highly unlikely due to the sensitive nature of the data. However, there are significant amounts of data available from various organisations that can be mined for the purposes of research. Data held in large databases, such as government databases, are an effective and useful source of existing data for a range of disciplines. Using existing data for research assists in providing evidence for the monitoring of social and economic indicators and for the development of international, national and state policies.

REFERENCES

Castle J (2003). Maximizing research opportunities: secondary data analysis. J. Neurosci. Nursing, 35(2): 287.

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39: 11, 27-34.

Gorard S (2001). Quantitative methods in educational research: the role of numbers made easy. London: Continuum.

Holscheimer M (1994). Data Mining: the search for knowledge in databases. The Netherlands: Amsterdam.

Isaac S, Michael W (1997). Handbook in Research and Evaluation. San Diego, California: Educational and Industrial Testing Service.

Jones FL, McMillan J (2001). Scoring occupational categories for social research: a review of current practice, with Australian examples. Work, Employment Soc., 15(3): 539-563.

Knipe S (2009). Young people in education and employment: The data trail. Germany: LAP LAMBERT Academic Publishing AG and Co.

McMillan J, Beavis A, Jones FL (2009). The AUSEI06: A new socioeconomic index for Australia. Journal of Sociology, 45(2): 123-149.

Miller H, Han J (2000). Geographic data mining and knowledge discovery: an overview. Retrieved July 25, 2009 from http://www.geog.utah.edu/~hmiller/papers/GKD.

Oshima TC, Domaleski C (2006). Academic performance gap between summer -birthday and fall-birthday children in grades k-8 [Electronic Version]. J. Educ. Res., 99: 212-217.

Rew L, Koniak-Griffin D, Lewis M, Miles M O'Sullivan A (2000). Secondary data analysis: new perspective for adolescent research. Nursing Outlook 48 223-9.

Schumacher S, Mc Millan J H (1993). Research in education a conceptual introduction (3rd ed.). New York: Harper Collins College Publications.

Wiersma W, Jurs S. (2005). Research methods in education: an introduction (8th ed.). Boston: Pearson/Allyn and Bacon.