The Australian Economic Review, vol. 40, no. 2, pp. 208–15

*Data Surveys*

# The Expanded Cross-National Equivalent File: HILDA Joins Its International Peers

Richard V. Burkhauser and Dean R. Lillard*
Department of Policy Analysis and Management
Cornell University

## 1. Introduction

Cross-national research using large representative datasets is still relatively new. It has only been over the last 25 years that harmonised cross-sectional data have become available to the international research community, primarily via the Luxembourg Income Study (LIS).[1] The LIS data allow researchers to measure the effects of alternative social policies and to compare socio-economic characteristics of different countries' populations from a cross-sectional perspective.

It is only in the last decade that several ex post harmonised country panel datasets have become widely available. Yet these panel studies have already been recognised as important sources of data that analysts need in order to investigate more fully levels of, and differences in, the relative economic well-being of the residents of different countries. Researchers are using panel data to analyse cross-national patterns of income mobility, poverty dynamics, and employment duration. They are also increasingly turning to panel data to compare and analyse cross-national differences in, and determinants of, individual and population health.

* The Cross-National Equivalent File (CNEF) has been funded over the years by the United States National Institute on Aging, the German Institute for Economic Research and Cornell University. This project is a collaborative effort with researchers in our CNEF partner institutions.

As principal investigators of the most widely used ex post harmonised country panel dataset—the Cross-National Equivalent File (CNEF)—and as researchers interested in comparing outcomes across countries, we were eager to bring the most important new nationally representative panel—the Household, Income and Labour Dynamics in Australia (HILDA) Survey—into the CNEF. In partnership with the Australian Department of Families, Community Services and Indigenous Affairs and the Melbourne Institute of Applied Economic and Social Research, we are delighted to report that the HILDA Survey is part of the 2006 CNEF data release. In this article we introduce Australian readers to the CNEF and its potential for researchers interested in comparing economic well-being, labour market and health outcomes in Australia with those in Canada, Germany, Great Britain and the United States.

## 2. Genesis and Evolution of the CNEF

Even the most sophisticated national surveys are unlikely to have cross-national comparability as a survey goal. Hence, while most national surveys use equivalent measures of age and gender, there is no international standard for measuring complex concepts like income, education, health or employment. Thus, researchers interested in doing cross-national work must investigate the institutions, laws and cultural patterns of a country in order to ensure that the variables they create for their analyses are equivalently defined across countries.

To reduce part of this burden, the LIS was developed. It brought together nationally representative micro-level household survey data from over 25 countries and attempted to make them comparable (see deTombeur et al. 1994). This innovative standardisation project greatly enhanced the ability of researchers to conduct cross-national comparative studies. However, it has two major limitations. First, the LIS is based primarily on restricted access national data sources. Researchers wishing to use the LIS data must access it through Luxembourg and must accept all LIS standardisation rules without access to the original data sources. Second, the data are cross-sectional and, hence, cannot be used for dynamic analysis.

In contrast, while the CNEF builds on the LIS model, it does so using multiple waves of longitudinal data from Canada, Germany, Great Britain, the United States and now Australia. What distinguishes the CNEF from other data harmonisation efforts is that the development of the data is driven by research questions. All variables flow from the work of experienced researchers who have developed cross-nationally these comparable measures for their own analyses. Thus, the standardised dataset we maintain is an amalgam of the knowledge of many researchers answering a diverse set of questions.

## 3. An Overview of the CNEF 1980–2005

Researchers at Cornell University along with colleagues from the Institute for Social and Economic Research at the University of Essex, the German Institute for Economic Research (Deutsches Institut für Wirtschaftsforschung—DIW) in Berlin, Statistics Canada in Ottawa, the Survey Research Center at the University of Michigan, and the Melbourne Institute of Applied Economic and Social Research at the University of Melbourne have developed and tested algorithms that place information from five panel surveys into a framework of comparably defined variables for use in cross-national research. These five panel surveys are the British Household Panel Survey (BHPS), the German Socio-Economic Panel (GSOEP), the Canadian Survey of Labour and Income Dynamics (SLID), the United States Panel Study of Income Dynamics (PSID) and the HILDA Survey. Using these panel surveys, researchers created a longitudinal micro-database known as the Cross-National Equivalent File 1980–2005 (CNEF). The CNEF is administered at Cornell University.

The CNEF was created to increase the accessibility and use of panel data among cross-national researchers and to assist current users of each panel survey in the creation of comparably defined cross-national variables. The CNEF is a dynamic and evolving body of data that unites comparably defined variables from these surveys into a single data file that can be used independently of, or in tandem with, the original survey data. It is designed to allow cross-national researchers not experienced in panel data analysis to access simplified versions of these panels, while providing experienced panel data users comparable variables that can be easily merged with data from the original parent surveys.

The 2006 CNEF release contains data from 1980 to 2005. The data include standard demographic information, household income and its components, and individual data on employment, labour earnings and health. Also included are cross-sectional and longitudinal sample weights, and macroeconomic indicators for each country. The CNEF is updated each year as additional waves of its five panels become available.

## 4. CNEF Country Data Sources

Each country survey collects similar information on household composition, income, employment, housing, health and demographic characteristics. In addition, each survey collects information about a variety of personal attributes, opinions and life choices. However, there are differences between the surveys. While each survey contains a core set of economic variables, each year some variables are added and deleted and some questions are reworded. The consequence is that some variables may be less comparable or sometimes are no longer comparable across surveys or within surveys over time.

Furthermore, the country datasets differ in their collection methods. The PSID differs in two distinct ways from the other panel studies. While all five panel studies collect information annually, since 1997 the PSID has collected data every other year. Further, in the PSID, interviews are only conducted with the head of a household, with all information about additional household members collected from this interview. In contrast, the other surveys interview all household members aged 16 years and older.[2] The SLID differs from the other panel studies because it has a rotating sample design. It consists of overlapping samples, each of which is followed for six years; the last three years of the older panel overlap with the first three years of the newer panel. Where possible, CNEF variables are created to be equivalent across surveys and over time. Where not possible, the differences are noted in the documentation so that researchers can take them into account in their analyses.

### *4.1 The BHPS <http://www.iser.essex.ac.uk/ulsc/bhps/>*

The BHPS began in 1991 with a sample of just over 5500 households containing approximately 10000 individuals. Households were selected based on postal code of residence. The original BHPS sample represents the population of households with postal codes in England, Wales and Scotland. Several booster samples have been added. In wave 7 (1997), a sample of new households was added as part of the European Community Household Panel Study. In 1999, a booster sample of Scottish and Welsh households was added. Finally, in 2001 a sample of households from Northern Ireland was included. From 1994 to 2000 (waves 4–10), the BHPS separately interviewed youth aged 11–15. All current BHPS households contain at least one member who was either part of the original 5500 households, part of the first wave of booster sample households, or born to a member of one of these households. The same individuals are re-interviewed each year. Those who split off from their original households are followed and the entire new household is included in subsequent samples. Children in original households are interviewed once they reach age 16 and continue to be interviewed when they leave their original households. For a more complete discussion of BHPS data see Lynn (2006).

### *4.2 The GSOEP <http://www.diw.de/english/sop/index.html>*

The GSOEP is the English-language public-use version of the Socio-Economic Panel (SOEP), a longitudinal dataset begun in 1984. The SOEP began with a sample of 6000 households living in the western states of the Federal Republic of Germany, including a disproportionate number of non-German migrant workers.

In November 1990, the eastern states of Germany were reunited with the western states of the Federal Republic of Germany. In June 1990, the DIW began a survey of families in the eastern states and merged these data with the existing SOEP population to provide a representative sample of reunited Germany. In 1994 and 1995, two new samples of households were added. Five hundred and twenty-two households were added that included at least one household member who had immigrated into the western states of Germany after 1984. In 1998 the sample was 'refreshed' with 1067 households randomly selected from the population of private households in Germany. In 2000 another 6052 households were added in this manner. Finally, in 2002, 1224 households were chosen from the population of households with monthly income of at least 3835 euros. In the second wave, the high income households in this sample were followed only if their monthly income was at least 4500 euros. The SOEP includes weights for each sample to allow researchers to generate sample statistics that represent the population of reunited Germany.

The Department of Policy Analysis and Management at Cornell University provides the GSOEP to non-European Union researchers. For confidentiality reasons the GSOEP is a 95 per cent sample of the full SOEP. We use the GSOEP in the CNEF. For a more complete discussion of the GSOEP, see Wagner,

Burkhauser and Behringer (1993). For a more complete discussion of the SOEP, see Haisken-DeNew and Frick (1998).

### *4.3 The PSID <http://psidonline.isr.umich.edu/>*

The PSID began in 1968 with a sample of 5000 families, representing a disproportionate number of low income individuals. All current PSID families contain at least one member who was either part of one of the original families or born to a member of one of them. Although the original sampling scheme disproportionately selected individuals from low income families, a representative sample of the US population can be obtained by excluding the original oversample from the data or by applying the sample weights provided with the data. Starting in 1997 the PSID began administering its survey every other year and no longer following every member or related member of families in the low-income oversample population. For a more complete discussion of the PSID, see Hill (1992).

### *4.4 The SLID <http://www.statcan.ca/start.html>*

The SLID began in 1993 with a sample of about 15000 households, containing approximately 30000 adults. The SLID survey differs from the other surveys in that each of its panels lasts only six years. In part, the limited length of these panels was chosen to keep the sample population representative of the national population. In 1995, three years after the first panel was surveyed, a second six-year panel was started. The same procedure was followed thereafter. This three-year overlap was chosen to maintain continuity in the data. As in the other surveys, all current SLID families contain at least one member who was part of, or born to one of, the initial household samples.

### *4.5 The HILDA Survey <http://melbourneinstitute.com/hilda/>*

The HILDA Survey is a household-based panel study, begun in 2001, that collects information about economic and subjective well-being, labour market dynamics, and family dynamics of Australian households. Wave 1 includes 7682 households and 19914 individuals. Interviews are conducted annually with all adult members of each household (Watson and Wooden 2004). HILDA panel members are followed over time in the same way as in the other CNEF samples. Each January a new wave of HILDA data is released and the wave is immediately incorporated into subsequent CNEF data updates.

## 5. The 2006 CNEF Release

The 2006 CNEF release includes subsets of the five original country panels (see <http://www.human.cornell.edu/che/PAM/Research/Centers-Programs/German-Panel/cnef.cfm>). Table 1 reports the status of each year of country data at the time of the 2006 CNEF release. Years marked P denote country data that have been collected but were not available in time for the 2006 CNEF release. The 2005 HILDA data were publicly released in January 2007 and are now being incorporated into the CNEF. Access instructions are provided in the CNEF documentation.

The 2006 CNEF release includes almost 100 variables covering topics such as demographics (age, sex, marital status etc.), employment (annual work hours, occupation, industry etc.), household composition, household and individual income (household income before and after government taxes and transfers, earnings, and income from various sources), geographic residence (state or province), health (see below), person and household identifiers, and individual and household sampling weights.

Variables have identical names, labels and, in most cases, value formats. The variable's name reflects the variable's content—the first letter of the variable name represents the variable's category: demographic (D), employment (E), household composition (H), income (I), weighting (W), sample identifiers (X), location (L) and medical or health (M). The last four digits of each variable name indicate the survey year from which the variable was drawn. This parallel structure allows researchers to use the same computer programs to

**Table 1 Years of Data Included in the 2006 CNEF Release by Country Dataset**

| | *HILDA* | *SLID* | *BHPS* | *GSOEP* | *PSID* |
|---|---|---|---|---|---|
| 1980 | | | | | X |
| 1981 | | | | | X |
| 1982 | | | | | X |
| 1983 | | | | | X |
| 1984 | | | | X | X |
| 1985 | | | | X | X |
| 1986 | | | | X | X |
| 1987 | | | | X | X |
| 1988 | | | | X | X |
| 1989 | | | | X | X |
| 1990 | | | | X | X |
| 1991 | | | X | X | X |
| 1992 | | | X | X | X |
| 1993 | | X | X | X | X |
| 1994 | | X | X | X | X |
| 1995 | | X | X | X | X |
| 1996 | | X | X | X | X |
| 1997 | | X | X | X | X |
| 1998 | | X | X | X | |
| 1999 | | X | X | X | X |
| 2000 | | X | X | X | |
| 2001 | X | X | X | X | X |
| 2002 | X | X | X | X | |
| 2003 | X | X | X | X | X |
| 2004 | X | P | X | X | |
| 2005 | X | P | P | X | P |

*Note*: X denotes available; P denotes in preparation; and a blank denotes not collected.

analyse data from all panels. Researchers can look through descriptive statistics for all CNEF samples on the CNEF web page.

Additional variables will be added to the CNEF as expert researchers develop them in the course of their substantive research projects. In this way, CNEF variables are vetted by the scholarly community. Like all peer-reviewed research, our variables are subject to criticism and, if warranted, are subject to correction. To facilitate such corrections, each country's CNEF data files are accompanied by documentation that describes the algorithm that was used to construct each variable. The algorithm lists the original variable names so that researchers can test alternative constructions of the variable in question.

The 2006 CNEF release contains a set of recently added variables that measure health behaviour and health outcomes of BHPS, GSOEP and PSID respondents. Their addition was timely since the GSOEP and the PSID had added modules with detailed questions about health status and behaviour. The new health variables include data on past and present medical conditions ranging from asthma to cancer, as well as information on height and body weight. Height and weight will be included in the HILDA-CNEF when the 2006 wave of HILDA data is released in January 2008.

### *5.1 Using CNEF Data*

The CNEF provides cross-national researchers not experienced in panel data analysis with a simplified version of each country's panel. Thus, the CNEF is distributed as a stand-alone data source, independent of the original surveys. The data are stored as rectangular data files—one for every year of survey data for each country.

The CNEF includes the unique person and yearly household identifiers from the original surveys to allow users to merge CNEF data more easily with subfiles from the full country panels. Finally, in addition to providing comparably coded versions of existing survey variables, the CNEF contains a set of constructed variables that are not directly available in any of the original surveys. These variables include measures of household income before and after taxes, estimated household tax burdens, and household size-adjusted median income for the population. Many of these variables cannot be computed without significant effort on the part of individual researchers.

For example, CNEF files include a measure of total household income after taxes and transfers (post-government income). This variable is the sum of labour earnings, asset flows, private transfers, public transfers, imputed rental value of owner-occupied housing, and other income of all individuals in a given household minus federal income and payroll taxes. Except for the SLID (which already includes taxes

paid), researchers familiar with the tax structure in their country wrote programs to estimate the tax burdens of respondents' households in each CNEF parent survey. For the PSID we use the National Bureau of Economic Research tax simulation program written by Daniel Feenberg (see Feenberg and Coutts 1993). Elena Bardasi, Stephen Jenkins and John Rigg of the Institute for Social and Economic Research at the University of Essex wrote the program for the BHPS (Bardasi, Jenkins and Rigg 1999). Johannes Schwarze of Bamberg University wrote the program for the GSOEP (Schwarze 1995). Simon Freidin, Nicole Watson and Bruce Headey of the University of Melbourne wrote the program for the HILDA Survey (Goode and Watson 2006).

While post-government income and other created variables are not available in the original country panel surveys, they are available in the CNEF in a comparable format. Because each of the created CNEF variables is the product of ongoing or completed research, variables included in the data file can be traced back to the specific papers in which they were developed. One of the benefits of this approach is that when a researcher is considering whether or not to use a particular variable, he or she can refer to the cited papers for examples of how that variable was developed and used.

### *5.2 Recent CNEF-Based Research*

Recent papers using CNEF data investigate topics over a broad range: economic well-being of widows (Burkhauser et al. 2005), obesity (Cawley, Grabka and Lillard 2005), health and income inequality (Lillard and Burkhauser 2005), marital dissolution (Andre$\beta$ et al. 2006), statistical measurement of income distributions (Clementi, Gallegati and Kaniadakis 2007), pension policies (Giavazzi and McMahon 2007), income mobility (Jenkins and Van Kerm 2006), poverty (Valletta 2006) and happiness (Zimmermann and Easterlin 2006). Researchers have already begun using HILDA data in conjunction with CNEF data to compare income inequality in Australia with inequality in other CNEF countries (Leigh 2005). These published and unpublished cross-national comparative papers are a sampling of the wide range of studies made possible by the equivalised country data contained in the CNEF. The addition of the HILDA Survey further enhances the value of the CNEF for cross-national research.

## 6. How to Obtain CNEF Data

Because the original PSID data are publicly available, we are able to post PSID-CNEF files on our web site (see above) for public use. To access BHPS-CNEF, GSOEP-CNEF or HILDA-CNEF files you must first apply for and be approved to use these data by the respective country's data manager. Once approved, email or fax us the approval documentation and we will send you the CNEF CD. To access SLID-CNEF files you must first be a registered CNEF user. SLID-CNEF data are not included on the CNEF CD, but all registered CNEF users can submit their programs to Statistics Canada. Staff at Statistics Canada will run these programs and return log and output files that meet confidentiality requirements.

The one-time registration fee to become a CNEF user is US$125, payable to Cornell University. For greater detail on how to access these data, visit the CNEF web page at <http://www.human.cornell.edu/che/PAM/Research/Centers-Programs/German-Panel/cnef.cfm> or send an email message to <cnef@cornell.edu>.

*February 2007*

## Endnotes

1. See Burkhauser and Smeeding (2001) for a review of the creation of nationally representative micro data for domestic use and their harmonisation for cross-national policy research.

2. In the case of the HILDA Survey, all persons aged 15 years or older are interviewed.

## References

Andre$\beta$, H. J., Borgloh, B., Bröckel, M., Giesselmann, M. and Hummelsheim, D. 2006,

'The economic consequences of partnership dissolution—A comparative analysis of panel studies from Belgium, Germany, Great Britain, Italy, and Sweden', *European Sociological Review*, vol. 22, pp. 533–60.

Bardasi, E., Jenkins, S. P. and Rigg, J. A. 1999, 'Documentation for derived current and annual net household income variables, BHPS waves 1–7', Institute for Social and Economic Research Working Paper 99-25, University of Essex.

Burkhauser, R., Giles, P., Lillard, D. R. and Schwarze, J. 2005, 'Until death do us part: An analysis of the economic well-being of widows in four countries', *Journal of Gerontology*, vol. 60B, pp. S238–46.

Burkhauser, R. V. and Smeeding, T. M. 2001, 'The role of micro-level panel data in policy research', *Schmollers Jahrbuch: Journal of Applied Social Science Studies*, vol. 121, pp. 469–500.

Cawley, J., Grabka, M. and Lillard, D. R. 2005, 'A comparison of the relationship between obesity and earnings in the U.S. and Germany', *Schmollers Jahrbuch: Journal of Applied Social Science Studies*, vol. 125, pp. 119–29.

Clementi, F., Gallegati, M. and Kaniadakis, G. 2007, 'Kappa-generalized statistics in personal income distribution', *European Physical Journal B*, forthcoming, E-print: arXiv:physics/0607293.

deTombeur, C., Milne, D., Warner, U., Gornick. J. and Randell, R. 1994, 'LIS information guide', Luxembourg Income Study Working Paper no. 7, Walferdange, Luxembourg.

Feenberg, D. and Coutts, E. 1993, 'An introduction to the TAXSIM model', *Journal of Policy Analysis and Management*, vol. 12, pp. 189–94.

Giavazzi, F. and McMahon, M. 2007, 'Anticipating reforms: Saving and work in Germany', unpublished paper, London School of Economics.

Goode, A. and Watson, N. (eds) 2006, *HILDA User Manual – Release 4.0*, Melbourne Institute of Applied Economic and Social Research, University of Melbourne, viewed January 2007, <http://www.melbourneinstitute.com/hilda/doc/doc_hildamanual.htm>.

Haisken-DeNew, J. and Frick, J. (eds) 1998, *DTC: Desktop Companion to the German Socio-Economic Panel Study (GSOEP), Version 2.2*, German Institute for Economic Research, Berlin, viewed February 2007, <http://www.diw.de/english/sop/service/dtc/index.html>.

Hill, M. 1992, *The Panel Study of Income Dynamics: A User's Guide*, Sage Publications, Beverly Hills, California.

Jenkins, S. P. and Van Kerm, P. 2006, 'Trends in income inequality, pro-poor income growth, and income mobility', *Oxford Economic Papers*, vol. 58, pp. 531–48.

Leigh, A. 2005, 'Permanent income inequality: Australia, Britain, Germany, and the United States compared', paper presented to 2005 HILDA Survey Research Conference, University of Melbourne, 29–30 September.

Lillard, D. R. and Burkhauser, R. V. 2005, 'Income inequality and health: A cross-country analysis', *Schmollers Jahrbuch: Journal of Applied Social Science Studies*, vol. 125, pp. 109–18.

Lynn, P. (ed.) 2006, *Quality Profile: British Household Panel Survey: Version 2.0: Waves 1 to 13: 1991–2003*, Institute for Social and Economic Research, University of Essex, Colchester, viewed January 2007, <http://www.iser.essex.ac.uk/ulsc/bhps/quality-profiles/BHPS-QP-01-03-06-v2.pdf>.

Schwarze, J. 1995, 'Simulating the federal income and social security tax payments of German households using survey data', Cross-National Studies in Aging Program Project Paper no. 19, Center for Policy Research, The Maxwell School, Syracuse University.

Valletta, R. G. 2006, 'The ins and outs of poverty in advanced economies: Government policy and poverty dynamics in Canada, Germany, Great Britain, and the United States', *Review of Income and Wealth*, vol. 52, pp. 261–84.

Wagner, G. G., Burkhauser, R. V. and Behringer, F. 1993, 'The English language public

use file of the German Socio-Economic Panel', *Journal of Human Resources*, vol. 28, pp. 429–33.

Watson, N. and Wooden, M. 2004, 'The HILDA Survey four years on', *Australian Economic Review*, vol. 37, pp. 343–9.

Zimmermann, A. C. and Easterlin, R. A. 2006, 'Happily ever after? Cohabitation, marriage, divorce, and happiness in Germany', *Population and Development Review*, vol. 32, pp. 511–28.