The Measurement of Labour Productivity using the Annual Enterprise Survey and LEED

Results of a data integration and data quality assessment

Sylvia Dixon

Department of Labour
and Statistics New Zealand

May 2007

Acknowledgements and notes on the data

This work was undertaken while the author was on secondment to Statistics New Zealand. The research was funded by the Department of Labour. The project benefited greatly from input or comments from the following people: Neil Kelly, Tissa Abeykoon, Stuart Jones, Julie Smith, Penny Barber, Richard Fabling, Dave Maré, Dean Hyslop, Sarah Crichton, and Steven Stillman.

Any views expressed are those of the author and do not purport to represent those of Statistics NZ or the Department of Labour. Any errors are the sole responsibility of the author.

Access to the data used in this study was provided by Statistics NZ under conditions designed to give effect to the security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person or firm. The tables in this paper contain information about groups of people so that the confidentiality of individuals is protected.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act. These tax data must be used only for statistical purposes, and no individual information is published or disclosed in any other form, or provided back to Inland Revenue for administrative or regulatory purposes. Any discussion of data limitations or weaknesses is in the context of using the Linked Employer-Employee Database (LEED) for statistical purposes, and is not related to the ability of the data to support Inland Revenue's core operational requirements. Careful consideration has been given to the privacy, security and confidentiality issues associated with using tax data in this project. Any person who had access to the unit record data has certified that they have been shown, have read and have understood Section 87 (Privacy and Confidentiality) of the Tax Administration Act. A full discussion can be found in the LEED Project Privacy Impact Assessment paper (Statistics NZ, 2003).

Abstract

This paper reports the findings of a feasibility study in which data from the Annual Enterprise Survey (AES) were linked to data from the Linked Employer-Employee Database (LEED) and enterprise-level measures of labour productivity and financial performance were constructed. An assessment was then made of the strengths and weaknesses of these data for research purposes. The overall purpose of the exercise was to explore the benefits that may be gained from linking business data to LEED.

In the course of the project, many useful insights were gained into the strengths and weaknesses of both AES and LEED at unit record level. This paper summarises those insights for the benefit of researchers who are considering the use of AES microdata in their own research, with or without other data sources such as LEED.

The paper covers issues such as the size and representativeness of the AES sample; whether AES can provide a representative longitudinal sample for use in longitudinal analysis; the longitudinal correlation of AES responses; the quality of the match obtained between AES and LEED records; and whether labour productivity measures that are constructed using AES measures of value-added and LEED employment data are comparable with other firm-level labour productivity measures.

The findings of the investigation indicate that a longitudinally-linked AES-LEED dataset is complete enough and of sufficiently good quality to be used in exploring a class of research problems that require longitudinal enterprise data. However, there is measurement error in the data, caused by non-response in AES and other data collection limitations. Researchers need to be aware of the data quality issues that exist and take care when drawing inferences from the data.

Contents

1. Introduction

This paper reports the findings of a feasibility study in which data from the Annual Enterprise Survey (AES) were linked to data from the Linked Employer-Employee Database (LEED), and enterprise-level measures of labour productivity and financial performance were constructed. An assessment was made of the strengths and weaknesses of these data for research purposes. The overall purpose of the exercise was to explore the potential benefits of linking business data to LEED.

In the course of the project, many insights were gained into the quality and strengths of weaknesses of both AES and LEED at unit record level. These insights are summarised in this paper for the benefit of other researchers who wish to know more about the quality of AES or LEED unit record data and its potential for research. (Note, however, that access to unit record data from both AES and LEED is restricted for confidentiality and other reasons. Researchers should consider the relevant access conditions carefully before planning a project.[1])

AES collects detailed measures of activity and financial performance from a representative sample of firms each year. It collects or derives measures of sales, income, expenses, gross output, value-added, and profits. AES also measures assets and liabilities (at balance sheet values) and net additions to capital stock during the year. LEED contains detailed workforce data, including measures of employment, earnings, and workforce characteristics, for all employees in the economy.

By linking AES and LEED data at enterprise level, it is possible to construct firm-level measures of labour productivity and profitability per worker. In addition, LEED provides information on the identity of every employee who has worked for a firm from 1999 to the present, which means that summary measures of employee attributes or activities can be accurately constructed at establishment or enterprise level, and linked to other enterprise records. Examples of workforce data that could in principle be linked to enterprise records using LEED identifiers include age, ethnicity, educational attainment, geographical location, turnover and participation in training. Some of these variables are already available in LEED, while others may become available in the medium-term future.

A linked AES-LEED dataset could potentially provide a foundation for any research that requires firm-level measures of productivity or financial performance, linked to workforce data. For example, given appropriate supplementary data, one might
use linked AES-LEED data to investigate questions such as:

- whether employee turnover rates, remuneration patterns, or other aspects of firms' human resource practices, are systematically related to the productivity of firms or their productivity outcomes
- how human resource practices change over the life cycle of the firm
- what types of firms invest in industry training (least productive or most productive)

1 Access conditions differ for the two data sources. Academic and government researchers can apply for permission to use unit record AES data in the Statistics NZ datalab. Depending on who is sponsoring the research, a number of conditions may have to be met, such as demonstrating that the research is of significant public interest. See http://www.stats.govt.nz/products-and-services/datalab.htm for more information on the criteria and process of applying to do research in the datalab. Access to the microdata that are held in the Linked Employer-Employee Database is more restricted, reflecting the strict confidentiality requirements of the Tax Administration Act. Access conditions are governed by the LEED Service Agreement between Statistics NZ and Inland Revenue. The data can only be used by approved Statistics NZ employees (including researchers seconded to Statistics NZ and approved by Inland Revenue) and cannot be accessed through the datalab.

- whether the type or amount of industry training undertaken has an effect on firm performance
- the effects of increased educational levels on productivity
- whether improvements in productivity lead to earnings increases.

Linked AES-LEED data could also provide useful measures of firm performance for integration into the LEED database. This type of data could enhance any LEED-based study in which employer behaviour or outcomes are modelled.

While every research question has its own unique data requirements, this project attempted to investigate data quality issues that are likely to be common across many research projects.

A linked dataset was created for the project and its properties were investigated. The analysis was structured around the following questions:

1. What are the strengths and weaknesses of AES for measuring output and value-added at firm level?
2. What are the strengths and weaknesses of LEED for measuring firm-level labour inputs?
3. How well can AES and LEED records be linked, and what is the quality of the match?
4. Does a linked AES-LEED sample constitute a representative sample of economically-significant enterprises?
5. Are enterprise labour productivity measures constructed using linked AES-LEED records consistent with labour productivity measures constructed from other data sources?
6. Do firms get selected for AES and respond regularly enough to provide a reasonably complete longitudinal dataset?
7. Do firm responses to AES show a high degree of correlation from year to year, or do they show a high degree of variation (suggesting there may be significant measurement error)?
8. When analysed at firm level, are factor input, output and financial performance variables related to each other in the expected ways, both at a point-in-time and through time?

The population of study was all privately owned profit-making firms that had at least 0.5 employees and total employment (including working proprietors) of at least one person on average during the year. Because AES is a sample survey, we matched LEED data to the records of AES respondents and analysed the matched data using AES sampling weights, which are designed to weight firms so that they represent all economically-significant enterprises in New Zealand. The AES years 2000 to 2004 were used in this project because they overlap with LEED.[2]

Measures of labour productivity and profitability per person were constructed using AES-sourced measures of value-added and profits, and LEED-sourced measures of average monthly employment. To extend our measures of labour inputs, we also used LEED to derive measures of average earnings, the average age of the workforce and the gender balance of the workforce, for each enterprise. In addition, we used the employee fixed effect estimates recently constructed by Maré and Hyslop (2006) to derive a proxy measure of the average

2 AES 2005 has been released and could be used in future work.

skill of the workforce at each enterprise. This ‘experimental’ AES-LEED dataset was then used to investigate the questions above.

The working paper is structured in the following way. Section 2 provides background information on AES and LEED and an initial discussion of their strengths and weaknesses. Section 3 outlines the methods that were used in this study to link the two data sources and construct measures of labour productivity and other variables. Section 4 addresses the research questions listed above and contains the main results of the paper.

The main findings of the paper are summarised in Section 5. Section 6 provides an overall conclusion on the potential benefits of linking AES to LEED. Appendix A provides some further information on data quality issues that were encountered in the course of the project.

2. Data sources and their strengths and weaknesses for labour productivity research

2.1 Previous research

In recent years, a small but growing number of New Zealand researchers have gained access to unit record data from official business surveys for research purposes.

Maré and Timmins (2006), Law and McLellan (2005) and Law *et al* (2006) have used unit record data to construct enterprise-level labour productivity measures. Specifically, they linked data on firms’ GST-reported purchases and sales from the Business Activity Indicators (BAI) database to Business Demography data on their employment at February of each year, and constructed measures of output per person or output per hour. The papers by Law *et al* focus on the relationships between labour productivity and firm births, deaths, expansions and contractions. The paper by Maré and Timmins focuses on the effects of geographic concentration and agglomeration on labour productivity.

A key issue for firm-level productivity research is finding a good measure of value-added. The BAI dataset on total GST-reported purchases and sales has the advantage of providing data for the vast majority of economically-significant firms in the economy.[3] An important disadvantage is that the difference between taxable sales and taxable purchases is not necessarily a good measure of value-added. Some of the limitations of the GST-based net sales variable as a proxy measure of value-added are noted in Law *et al* (2006, p8) and Maré (2006, p19). Appendix B provides additional information on the differences between the accounting concept of value-added and the BAI net sales measure.

A project team within Statistics NZ is currently assessing the potential for greater use to be made of tax-sourced data in longitudinal business statistics and research. A substantial proportion of firms that are not in the AES sample supply data on their financial performance and financial position to Inland Revenue, through their IR10 (accounts information) returns. The project is looking at the potential for using these and other administrative data more effectively, so that financial data is available for a higher proportion of all businesses than at present. One possible outcome of the project is that economic researchers will have access to a wider range of financial data sources in the medium-term future.

3 Specifically, it covers all businesses that have registered for GST collection or to claim GST refunds. All businesses that conduct taxable activity are required to register for GST if their annual turnover is greater than $40,000.

2.2 The Annual Enterprise Survey

AES collects data on financial performance and financial position from a sample of all economically-significant[4] enterprises that operate in the New Zealand economy.[5] It is used to produce industry and national statistics on levels and changes in the financial performance, assets and equity of firms, including measures such as operating surplus, the return on assets, the return on equity, debt ratios and gearing. It also provides industry-level measures of economic activity that are used in the construction of the National Accounts. In 2004, the estimated population size was approximately 400,000 enterprises. The AES target population covers approximately 90 percent of New Zealand's GDP.

In 2004, around 20,900 organisations were directly surveyed through a postal survey; 240,000 were included in the sample through an analysis of data supplied to Inland Revenue on IR10 (accounts information) forms; and around 4,000 publicly-owned organisations were included using data supplied directly by central or local government.[6]

The postal survey is used to collect relatively detailed data from limited liability companies, branches of overseas companies, and non-profit organisations. Individual proprietors and partnerships are included in AES through the analysis of tax (IR10) data. In addition, all agricultural industries, with the exception of services to agriculture (A021), are included in AES solely through the analysis of tax data. A full-coverage approach is used in the tax component of the survey – all businesses that complete an IR10 and provide usable responses are included in the sample.[7]

AES was not designed to be a longitudinal survey. It was designed to produce annual estimates. However, firms that were selected into the AES postal sample at the last redesign (1999) have a high probability of being reselected each year if they do not change their legal form or industry and remain reasonably stable in size.[8] The full-coverage nature of the tax sample also means that continuing firms have a high probability of appearing in the AES dataset in successive years.

The survey collects data from firms with financial year end dates ranging from 1 October in the previous year, through to 30 September in the survey year. Each firm reports data for its own financial year, in current prices. There is no price synchronisation of the unit record data by Statistics NZ (although price adjustments are made when the data are used in the National Accounts).

4 An economically significant enterprise is defined as an enterprise which meets at least one of the following criteria: has greater than $40,000 annual GST expenses or sales; has more than two full-time equivalent paid employees; is in a GST-exempt industry except residential property leasing and rental; is part of a group of enterprises; is a new GST registration that is compulsory, special or forced (this means the business is expected to exceed the $40,000 boundary); is registered for GST and is involved in agriculture or forestry.

5 A small number of industries are excluded. These are residential property operators, foreign government representatives, religious organisations, and private households employing staff.

6 Central government data comes from the Crown Financial Information System in Treasury and local government financial data is compiled by Statistics NZ. Data on the insurance industries is obtained from the Government Actuary.

7 Responses are imputed for firms that have not supplied usable data. Consequently, firms in the tax component of the sample almost always have a weight of '1'.

8 All enterprises on the Business Frame are assigned a 4-digit random number between 0 and 1. In order to reduce respondent burden, each business survey administered by Statistics NZ selects its sample from within a certain zone of the random number line which does not overlap with the zones used by other business surveys. When the AES sample is redrawn each year, the same portion of the random number line is used, which means that most enterprises are reselected. The sample is stratified by firm size and industry, and sampling rates vary across strata. A firm's probability of reselection for AES will depend on its size and industry, among other factors.

Relevant variables in the AES data set include gross output, value-added, operating surplus, profit, total income, total expenditure, various measures of sales, salaries and wages paid to employees, compensation paid to self-employed persons, purchases and other operating expenses. Asset measures include a balance-sheet measure of fixed tangible assets; the value of net additions to plant, machinery and equipment during the year; and the value of net additions to land and buildings during the year.[9]

The AES data set also includes data on firm demographics, which is largely taken from the Business Frame at the time the sample is drawn. There are variables for business type (the legal form of the organisation), industry, institutional sector, the enterprise number of any corporate group the enterprise is part of, and the firm's GST sales and employee count at the time the sample was drawn.

2.3 Advantages and disadvantages of AES as a source of output or value-added data

AES uses detailed, industry-tailored questionnaires to measure output and value-added. The measurement approach is broadly consistent with the national accounting concept of value-added.

Because the purpose of AES is to generate aggregate statistics, not to produce a clean unit record dataset for research purposes, 'significance editing' is applied when the data are compiled. Data editing and checking focuses on the responses of firms that have large financial values and are likely to have a material impact on aggregate results. Less editing is done for small firms. This means that the quality of the data, at unit record level, may be uneven.

A disadvantage of AES for some research purposes is that it gathers data from a stratified sample of limited liability companies only, not the entire population. Although a high proportion of the largest firms in the economy are included, the sampling fractions for small and medium-sized companies are much lower. Most small and medium-sized limited liability companies are *not* included in the AES sample. Weights must be used to derive population estimates. The current weights were designed for a specific purpose (to give good industry-level and national estimates of firms' financial performance and financial position), and they may not be ideal for all research purposes.

The AES sample was not specifically designed to track enterprises over time. However, a high proportion of enterprises in the both the postal and tax samples are reselected each year. This means researchers can link enterprise responses across years (using firms' unique enterprise numbers), and construct longitudinal samples. While longitudinal analysis is certainly possible, panel continuity is reduced by attrition from the sample, by non-response, and by partial non-response, all of which are which are reasonably common. Further information on rates of attrition and non-response is provided in Section 4 below.

For some research purposes, the unit of collection could be a disadvantage. Some human resource practices (such as recruitment) tend to operate at the establishment level and may vary between establishments within the same enterprise. The association between human resource practices and productivity may be weaker for multi-establishment firms when productivity is measured at the enterprise level.

2.4 The Linked Employer-Employee Database

LEED is an administrative database containing monthly employment and taxable earnings data for all employees in the economy, linked to information on their employers. The core

9 See the Statistics NZ website www.stats.govt.nz for further details on AES variables.

data in LEED is derived from the Employer Monthly Schedule (EMS), a report that all firms with employees must submit to Inland Revenue each month when they make PAYE deductions. EMS lists all employees who worked at the firm in the last month, the amount of income they received, and the amount of tax that was deducted at source.[10]

Currently, all such records of taxable earnings from April 1999 onwards are held within the LEED database. Both individuals and employers in the tax system have unique identifiers which enable their records to be linked longitudinally through time.

We use a version of the LEED data in which EMS returns are allocated to geographic units and enterprises on the Longitudinal Business Frame (LBF). The LBF is a version of the Business Frame in which unique longitudinal identifiers have been assigned to business activity units (ie establishments), so that they can be accurately followed through time. The LEED-LBF structure was created to provide a longitudinal view of the labour market.[11]

In addition to the employment and earnings data, the LBF contains some additional information on establishment and enterprise characteristics, such as business type, institutional sector, and the date when the firm was first birthed on to the Business Frame.

LEED has information on each employee’s gender, age, and geographical area of residence, obtained from Inland Revenue sources.[12] Currently LEED does not have any measures of hours worked, or data on employees’ educational levels or occupations.[13]

At the time this study began, LEED did not include data on the numbers or incomes of self-employed people. For this study, data on the number of working proprietors employed at each firm were taken from the LBF. Statistics NZ actively maintained a Business Frame variable on the number of working proprietors employed at each establishment up until the middle of 2003, when there was a major change in Business Frame maintenance procedures. A decision was taken not to maintain the working proprietor field in the new system. Since then, no updates have been made to the Business Frame working proprietor counts for continuing enterprises.

During 2006, Statistics NZ constructed annual counts of self-employed persons and self-employed jobs using a combination of income tax sources. LEED statistics on self-employment were published for the first time in October 2006. Because the new method of identifying people who receive self-employment income from an enterprise is conceptually different from the old approach to counting working proprietors, it is likely to lead to different counts at enterprise level. The implications of this change for the task of estimating self-employed labour inputs is currently rather unclear and merits further investigation.

2.5 Advantages and disadvantages of LEED as a source of labour input data

LEED is currently the best available source of data on employees. An important advantage of LEED, relative to the Business Frame and Business Demography employee count, is that it

10 Two types of recipient are covered by EMS: those who have pay-as-you-earn (PAYE) tax deducted, who are employees, and those who pay withholding tax, who are a subset of self-employed individuals. In this study we discard the withholding tax records.

11 A small number of employers who were not recorded on the Business Frame have been identified in LEED and assigned unique longitudinal identity numbers.

12 The age, gender and residential location fields are imputed when actual data are not available from Inland Revenue.

13 A proxy measure of employee skill has been estimated in a separate LEED research project (Maré and Hyslop, 2006). For each worker, a ‘fixed effect’ has been calculated, providing an estimate of the portable earnings premium that they received, whichever firm they worked in during the period 1999–2005. See section 3.4.

provides monthly observations. This provides better information on the variation in labour inputs during the year, and better information on the labour inputs of firms that may have opened or closed during the year. For enterprises that are part of groups that make joint PAYE returns, the final LEED employee counts are also likely to be more accurate, because the employees of these groups have been allocated by Statistics NZ to the establishments and enterprises that belong to the group (using employee residence details and other data).

LEED is only semi-longitudinal at the enterprise level. This is because the LBF was designed to track establishments and repair 'false' births and deaths at establishment level (where an establishment is a geographical business unit). If an enterprise changes its legal structure or is sold to another owner it is normally assigned a new Inland Revenue number and a new enterprise number on the Business Frame, even if the fundamentals of the business are unchanged. The LBF creates consistent links between establishments that are believed to be the same establishment, and assigns them the same ID number, but firms' enterprise numbers are not revised or repaired. In addition, the current LBF continuity rules embody various compromises and do not always track establishments accurately (see Fabling, 2006).[14]

As noted above, working proprietor measures are in transition at present. At the time this analysis was undertaken there was no up-to-date or completely comprehensive source of data on the working proprietor labour inputs of enterprises. Somewhat outdated historical data were used in this analysis.[15] A different approach to measuring working proprietor employment will need to be adopted in future studies.

As noted above, both LEED and AES lack any data on the hours worked by employees and working proprietors. This is a significant limitation for the measurement of labour productivity. Researchers must either use a head count measure of labour inputs,[16] or impute hours worked in some way.

3. Method of constructing an experimental linked AES-LEED dataset

3.1 Introduction

This section outlines the methods that were used to link LEED and AES records and create measures of firm performance for the purpose of this study.

Section 3.2 describes the population of study. Section 3.3 describes the methods used to link AES and LEED. Section 3.4 describes the output and input measures in more detail and the method of constructing derived variables. Section 3.5 discusses some variable quality issues and the sample exclusions that were made in this study.

3.2 Population of study

The population of study is all privately owned, profit-oriented firms that had at least 0.5 employees per month, and total employment (including working proprietors) of at least 1.0

14 A project within Statistics NZ is looking at the options for repairing some of the 'false' breaks in enterprise number continuity on the Business Frame.
15 To develop that approach, more information is needed on the differences between historical working proprietor counts and the counts yielded by the recently-adopted strategy of using income tax returns to identify people who received self-employment income from enterprises.
16 It is possible to adjust the head count measure using LEED data on employees' monthly earnings. Maré and Hyslop (2006) calculate a full-time equivalent employment measure in which workers with earnings below a certain threshold, or concurrent income from a benefit, are assumed to be part-time employees, and are assumed to work fewer hours.

worker per month, on average during the year.[17] We restrict our focus to firms with average employment of at least one worker because labour productivity is a key variable of interest. In addition, because of the problems with the working proprietor data (discussed above) we decided to exclude working proprietor-only firms.

We use the Statistics NZ 'business type' classification to determine which firms are private profit-oriented businesses. We retain enterprises with a business type of 1 to 6, covering sole proprietorships, partnerships, limited liability companies, cooperatives, joint ventures, and branches of overseas companies.

Because AES is a sample survey, we match LEED data to the records of AES respondents and analyse the matched data using AES sampling weights, which are designed to weight firms so that they represent all enterprises in the AES population (see section 2.2). Specifically, the weights are calculated so that the weighted results accurately represent firms in each target industry (for financial performance variables) and each institutional sector (for financial position variables). The AES years 2000 to 2004 were used in this project because they overlap with LEED. The coverage of LEED begins in April 1999.[18]

3.3 Method of linking LEED data to AES records

The unit of data collection in AES is the kind of activity unit or KAU, which is an accounting unit within an enterprise. The vast majority of enterprises in AES have only one KAU, so there is a one-to-one match between the two. However, a small number have more than one KAU: 178 in 2001, 173 in 2002, 172 in 2003, and 155 in 2004. If an enterprise is included in the AES sample, all its KAUs are surveyed.

AES records must be aggregated to one record per enterprise before they can be linked to other business datasets. In this study, the aggregation simply summed input and financial performance variables such as fixed assets, value-added, gross output, profits, and salaries and wages, across KAUs belonging to each multi-KAU enterprise. Note that if there are intra-company transfers, this simple aggregation of KAU records could give incorrect results (for example, by counting sales and purchases that are actually transfers of goods between accounting units within the enterprise).

Enterprise characteristics that do not normally vary within an enterprise, such as sector and business type, were taken from the record of the enterprise's largest KAU (ranked in terms of average employee numbers). Industry code normally differs between KAUs. The correct enterprise-level industry code was therefore taken from the LBF.

Following aggregation to enterprise level, AES and LEED records can be easily linked using Statistics NZ's unique Business Frame enterprise numbers.

AES has an annual structure, while LEED has a calendar month structure. Moreover, different firms in AES have different financial year start and end dates. In this study, we selected LEED data for the 12 months best matching the financial year of each firm in AES.[19]

Note that if financial year balance dates were recorded incorrectly in AES, or if firms reported their previous year's financial data because their accounts were not complete at the time they

17 Part-year operators are included if their employment levels during months of operation, when divided by 12, allowed them to reach the average monthly employment threshold of 1.

18 However, it was necessary to drop 8.3 percent of AES records from the 2000 year, comprising firms with balance dates before 30 March 2000, because a full year of LEED employment data is not available for these firms.

19 A small percentage of firms have a balance date that falls *within* a month, but usually these are very close to the start or end of a month.

responded to AES, or if firms reported employee earnings to Inland Revenue in the months before or after the payments were actually made and recorded in the firm's accounts, then the data match could be incorrect purely for timing reasons. Currently we have no hard information on whether these are important sources of error (although the data matching issues discussed below suggest they could be).

3.4 Variable construction

Output and performance measures in AES

This study made use of the value-added and profit measures that are constructed by Statistics NZ as part of the processing and analysis of AES.

Value-added represents gross output minus intermediate consumption. For most industries in AES, 'gross output' is calculated in the following way:[20]

> sales and other operating income (where the latter includes items such as management fees, professional fees, and income from renting or leasing building, plant and machinery)
>
> + own account capital formation (representing goods produced for by a firm for its own use, eg production of machinery and equipment for use in the production process)
>
> + adjustments for changes in stocks of work in progress, trading goods and finished goods.

'Intermediate consumption' is calculated as:

> purchases and other operating expenses (the value of the goods and services that were consumed as inputs by the process of production)
>
> + adjustments for changes in stocks of raw materials.

'Profit' is calculated in AES as total income minus total expenditure. 'Total income' includes income from the sale of goods and services, income from rents, leases, interest and dividends, government grants or subsidies and any other non-operating income. 'Total expenditure' includes purchases of goods, services and raw materials, salaries and wages, indirect taxes, depreciation, and all non-operating expenses.

Measures of labour quantity and quality

In this study the main labour input measure is the average number of persons employed per month, obtained by summing the LEED-sourced count of employees and the LBF-sourced count of working proprietors. LEED provides calendar-month counts of the total number of employees who worked at the firm during the month (the exact start and finish dates of each employee are not generally available). We calculate each firm's average monthly number of employees by dividing the sum of employee-months over the financial year by 12.

20 This is a simplification as there are further refinements and industry variations. For example, excise taxes are added to gross output in the manufacturing and wholesaling industries, and road user charges are added in transport. In the heath sector, 'sales and other income' includes certain types of medical income from the government. In financial services, the value of interest received minus interest paid is an important component of gross output. In insurance, the value of premiums received minus premiums paid is an important component.

Firms that did not operate for all of the year are included in AES if they operated for at least six months. Only about 4 percent of firms in the study sample were part-year operators. In AES, no adjustment is made to the gross output or value-added totals for part-year operators. For consistency, we do not adjust the LEED employee count for part-year operation. This means, for example, that a firm with one employee for six months of the year has a value of 0.5 for the 'mean employees' variable.

The data on working proprietors (taken from the LBF) provides annual employment counts only. If a firm reported in AES that it operated for fewer than 12 months, we adjust their working proprietor count using the 'months of operation' variable. For example, if a firm had one working proprietor and reported that it operated for 10 months, the adjusted working proprietor count is 10/12 = 0.83.

We construct an estimate of total annual hours worked as a secondary measure of labour inputs. Estimates of the average weekly hours that were worked in each industry, over the period from April 1999 to March 2004, were obtained from the Household Labour Force Survey (HLFS) and multiplied by the firm-specific employment counts to get an estimate of each firm's total annual hours.

The HLFS-sourced estimates of employee hours were calculated at 3-digit, 2-digit, or 1-digit level, depending on the size of the industry. Three-digit data were only used if the 3-digit industry group had a minimum average sample size of 50 persons per quarter; otherwise 2-digit estimates were assigned. Our estimates of the average weekly hours of working proprietors were calculated at 1-digit industry level using HLFS self-employment data. The weekly hours averages were calculated for March financial years. This matches the actual financial year of the AES firms in the study sample for around 74 percent of cases. Note that the HLFS hours-by-industry data are subject to sizeable sampling errors, and they show year-to-year variability which is likely to be partly measurement error.

For the analysis of variations in productivity across firms, there is little to be gained by measuring labour productivity in units of hours rather than persons, since there is no additional variation within industries with this imputation method. However, industry differences in hours worked will affect inter-industry comparisons. The estimates of hours worked may therefore be of some value for comparing labour productivity across industries.

LEED currently contains information on employees' age and gender, but few other demographic or skill characteristics. In a separate project, Maré and Hyslop (2006) have estimated employee fixed effects for all employees in New Zealand, using the method of Abowd, Creecy and Kramarz (2002). These worker fixed effect estimates represent the portable earnings premium that each worker receives, whichever firm he or she works in. We make use of them in this project in order to test hypotheses about the relationships between labour force quality and productivity.

Specifically, the worker fixed effect estimated by Maré and Hyslop captures the premium that an individual earns relative to other workers of the same age and gender, and net of any employer-specific earnings premium.[21] Because differences in hours worked are not included in the estimation process, the employee fixed effect estimated by Maré and Hyslop includes both differences in earning capacity (which allow some individuals to work longer hours than others) and differences in wage rates.

As part of the same project, Maré and Hyslop also calculated establishment-level measures of the average age of employees and the proportion of employees that were female. After

21 The model they use to estimate worker fixed effects includes controls for the earnings level of the current employer and controls for the average earnings of a person of the same age and gender.

aggregating from establishment to enterprise level, we link their employee fixed effect and workforce demography variables to our experimental AES-LEED dataset.[22]

Measures of capital inputs

AES was not designed to measure the capital stock of enterprises or the flow of capital services used. In this project, we use the AES balance-sheet measure of total fixed assets as the best available proxy measure of capital intensity. Because of depreciation and other factors, this variable is quite likely to underestimate or overestimate the true level of productive capital that is available to the enterprises in the sample.[23]

Converting nominal values to real values

Producer Price Index (PPI) output series were used to inflation-adjust gross output. PPI input series were used to inflation-adjust intermediate consumption. These two variables were then recombined to get a measure of real value-added.[24]

Because the PPI series are quarterly in frequency, we took the annual average of the four quarters of the index corresponding to each firm's financial year. No special adjustment was made for firms with part-year operations because it isn't possible to tell which quarters they operated in.

The PPI series are constructed at a mixture of 3-digit, 2-digit and 1-digit levels. We used the most detailed level available.[25] In general, more disaggregated price indexes are calculated for the primary and secondary sectors of the economy than for services.

No PPI output series are available for some industries such as health and education. For firms in these industries we used the all-industries index to convert nominal values to real.

The PPI output series were also used to inflation-adjust nominal *profits*. The *fixed assets* data were price adjusted using the Capital Goods Price Index series for all groups (excluding residential buildings).

Measures of the average annual gross earnings paid to employees were deflated using the Labour Cost Index (LCI). Specifically, we used the LCI series for salaries and wages in private sector firms. These LCI indexes are calculated at 2-digit ANZSIC level for manufacturing and 1-digit level for the rest of the economy. Unfortunately, the LCI only adopted standard ANZSIC groups in 2001, and the industry groups used previously are not easily reconciled. For this reason, the all-industries index was used to inflation-adjust earnings for the period from April 1999 to March 2001. Industry-level indexes were applied from June 2001 onwards.

Derived productivity and profitability measures

The following measures of firm performance were derived:

22 We take a weighted average of the estimates for each establishment that belonged to the enterprise in the middle month of its financial year. The weights for each establishment are based on its share of the enterprise's total full-time equivalent employee numbers.
23 Depreciation rules mean that the decline in the balance-sheet value of an asset is often faster than the decline in its economic life.
24 Because a summary measure of intermediate consumption was not available on the AES dataset, we simply took the difference between gross output and value-added.
25 Some unpublished 3-digit series were obtained from the Prices team in Statistics NZ.

Performance:

- Value-added per person = real annual value-added /average monthly employment. Expressed in March 2002 dollar values and units of $1,000.
- Profits per person = real annual profits /average monthly employment.
- Value-added per hour = real annual value-added / annual sum of hours worked. Expressed in $ per hour.
- Profits per hour = real annual profits / annual sum of hours worked.

Log versions of these performance variables:

- Log value-added per person = log VA – log mean monthly employment.
- Log value-added per hour = log VA – log annual hrs /1,000.
- Log profits per person = log profits – log mean monthly employment.
- Log profits per hour = log profits – log annual hrs /1,000.

3.5 Imputation, item non-response and sample exclusions

AES was designed to provide accurate measures of financial outcomes, assets and liabilities at industry, sector and national levels. It was not designed to measure the economic activity or financial performance of individual firms. Consequently, the unit record data has some features that could reduce the utility of the data for some research purposes. Appendix A gives some further information on data quality issues.

Imputation of records

Non-response is reasonably common. Between 2001 and 2004 about 32 percent of all private enterprise records (including 33 percent of postal sample records and 31 percent of tax sample records) were imputed by Statistics NZ because the unit either did not respond in the postal survey or failed to provide usable data on a questionnaire or IR10 return.[26] Appendix A describes imputation methods.

Smaller firms are more likely to have their records imputed than larger firms.[27] This means that any analysis that excludes imputed responses is likely to under-represent small enterprises. Further information on the profile of the AES sample used in this analysis, before and after the imputed records were excluded, is given in section 4.2 below.

Item non-response and negative values

Currently, responses of ‘0’ are not clearly distinguishable from non-response in the AES data. In general, a missing response for a particular item in a questionnaire or IR10 form is coded to zero. Given this, researchers using AES have little option but to treat all zero values as likely non-response, even though some may genuinely be responses of zero.

After excluding imputed records, about 3 percent of private-sector enterprise records in AES have a negative value of value-added.[28] About 2.0 percent have zero value-added. Around 22 percent of private-sector enterprises report negative profits, ie losses, and 3 percent have a zero value for the profit variable. These sample proportions are far lower if firms without employees are excluded and firms are weighted by their number of employees. However, consideration still needs to be given to the appropriate treatment of firms with negative or zero values for key variables.

---

26 Responses are put through a series of editing checks and if they fail (for example, because key items are missing or there is substantial inconsistency between items) the form is discarded and a complete response is imputed.

27 Respondent follow-up procedures target larger firms.

28 Sampling weights were used when calculating the numbers in this paragraph.

A small percentage of records in the private-sector AES sample, after excluding imputed records, have a zero in the fixed assets field. It is not clear whether these firms genuinely had no fixed assets or simply did not provide data.

Sample exclusions

Firms with imputed AES data are excluded from the analysis in this paper. Because firms with a zero entry for 'value-added' nearly always had zero values of gross output and sales as well, they were taken to be incomplete respondents and were also dropped from the sample used in this analysis.

Two to three percent of firms in the remaining study sample had negative values for the value-added variable. Firms with negative values were initially retained in the analysis sample, but were dropped from analyses in which variables were converted to logarithms. Logs were used in many of the analyses reported below to reduce the impact of outliers on the estimates obtained.

Firms with zero fixed assets were retained in the initial analysis but excluded from analyses in which variables were converted to logs.

4. Analysis of the linked AES-LEED data

4.1 Introduction

This section investigates the completeness and quality of the linked AES-LEED data. It is structured in the following way:

Section 4.1 analyses the match rate obtained and quality of the match when AES and LEED records are linked. Section 4.2 describes the profile of the matched study sample, and compares it with the entire AES sample of private, profit-oriented firms with employees.

Section 4.3 presents national and industry-level summary statistics calculated using the labour productivity and profit per person measures derived in this study. Section 4.4 compares our industry-level estimates of average labour productivity with estimates generated by Law *et al* (2006) and Maré and Timmins (2006) using a combination of BAI and Business Demography data.

Section 4.5 examines the pattern of repeat responses by firms in the AES sample. The key issue to assess is whether a usable longitudinal dataset can be extracted from AES.

Section 4.6 considers the extent to which the AES-LEED input and output measures are correlated across time at firm level. Section 4.7 looks at the relationships between input and output variables, using correlations, simple production function models and other simple regression models. These are estimated on a cross-sectional basis first and then re-estimated using the panel structure of the data.

Section 4.8 considers the question of whether firms whose LEED-sourced and AES-sourced salary and wage data match poorly, suggesting matching or measurement error, are systematically different from other firms. It shows how the results obtained in section 4.7 change when these firms are dropped from the sample or allowed to have different parameters.

4.2 The AES-LEED match rate and differences between the matched and unmatched populations

The study population in this project is all private, profit-oriented enterprises that had at least 0.5 employees and at least 1.0 employed persons (including working proprietors), on average over the year.[29] The first column of table 4.2.1 gives an estimate of the number of enterprises in LEED that met these business type and employment criteria, in the month of March each year from 2000 to 2004, using data currently held in the LBF. On average, about 128,000 enterprises met the selection criteria each year. Note that we are using a ‘snapshot’ estimate of the study population. This is intended simply to provide a reasonably good indication of the size and properties of the study population.[30]

The next three columns of the table show the numbers of firms that were also part of (a) the postal component of the AES sample, (b) the tax component of the AES sample, and (c) the entire AES sample. On average, nearly 60,000 firms in the estimated study population were also in the AES sample. This means that around 47 percent of eligible firms (as identified here) were in both data sources.

Around 9,000 to 10,000 were included via the postal sample and the remaining 50,000-odd were included via the tax sample. These two sample components have very different profiles, as illustrated in the second and third columns. The postal sample over-samples medium-sized and large organisations. Consequently the mean employment level, salary and wage expenditure and value-added totals are high for this sample component. The tax sample, by contrast, contains a high proportion of small units.

The industry distributions indicate that a high proportion of units in the tax sample are farms (agriculture industry), construction businesses, shops (retail industry), or firms in property and business services. The postal sample, on the other hand, is more oriented towards industries in which large firms tend to operate, such as manufacturing.

When these two sample components are combined, the AES sample as a whole has some unusual properties. Both very small and very large firms are over-represented, while firms in the 5–49 persons size group are under-represented. The survey sampling weights are, of course, designed to adjust the unweighted counts accordingly.

---

29 As outlined in section 3, we assume that working proprietors present in February worked for the entire year unless we have evidence of part-year operation from AES. Employee numbers are available on a monthly basis in LEED.

30 The estimated population numbers were similar when we selected a snapshot of firms meeting the study criteria in the month of June, rather than March. Note that the LBF files used for this exercise are not exactly the same as the BF files that were actually used to select the AES sample each year from 2000 to 2004.

Table 4.2.1

Profile of the linked population

	LEED-estimated total study population as at March (1)	LEED and AES postal sample	LEED and AES tax sample	LEED and AES total (2)	Ratio (2)/(1)
	N	*N*	*N*	*N*	
Average 2000 to 2004	127,780	9,730	49,790	59,530	*0.47*
2000	122,510	8,620	49,880	58,510	*0.48*
2001	124,120	9,960	51,690	61,650	*0.50*
2002	126,590	9,990	50,200	60,200	*0.48*
2003	130,910	10,020	48,860	58,880	*0.45*
2004	134,770	10,080	48,310	58,400	*0.43*
	Means	*Means*	*Means*	*Means*	
Monthly employees (LEED)	9.5	59.0	3.3	12.4	*1.31*
Monthly employment (LEED + LBF)	11.0	60.4	4.8	13.9	*1.27*
Monthly employees (AES)	…	61.8	3.3	13.3	…
Total annual payroll ($000) (LEED)	288.7	2,121.0	61.4	398.3	*1.38*
Total annual payroll ($000) (AES)	…	2,098.1	59.8	393.1	…
Value-added ($000) (AES)	…	4,460.7	151.0	855.7	…
Persons employed	%	%	%	%	
1<5	56.7	29.1	81.3	72.7	…
5<10	24.5	17.0	13.1	13.7	…
10<20	11.3	16.2	4.1	6.1	…
20<50	5.3	17.1	1.3	3.9	…
50<100	1.3	9.8	0.2	1.8	…
100+	1.0	10.8	0.0	1.8	…
Industry distribution	%	%	%	%	
Agriculture	19.0	3.0	41.4	35.1	…
Mining	0.2	1.7	0.0	0.3	…
Manufacturing	10.1	19.1	4.3	6.8	…
Electricity, gas & water	0.0	0.3	0.0	0.1	…
Construction	10.6	7.9	8.6	8.5	…
Wholesale trade	6.7	11.8	1.7	3.4	…
Retail trade	17.2	13.1	13.0	13.0	…
Accomodation, restaurants & cafes	6.2	2.9	5.1	4.7	…
Transport & storage	3.8	5.1	2.4	2.9	…
Communication	0.6	0.4	0.8	0.7	…
Finance & insurance	1.7	6.6	0.4	1.4	…
Business services	13.7	14.3	11.2	11.7	…
Government	0.0	0.0	0.0	0.0	…
Education	0.8	3.2	0.5	1.0	…
Health & community services	4.6	5.2	5.4	5.4	…
Recreational & cultural services	1.5	3.1	1.2	1.5	…
Personal & household services	3.3	2.3	4.1	3.8	…

Note: The data in this table are unweighted.

Symbol:
… not applicable

Table 4.2.2 gives data on non-matched enterprises and some other anomalous cases. Non-matched enterprises are those that met the study population criteria but were not sampled for the AES postal survey, or were not included in the tax sample for some reason. Non-matched enterprises are more likely to fall into the 5–50 person size groups. The number of non-matched enterprises also increased over the 2000 to 2004 period, probably because the size of the AES sample was fixed while the population of enterprises was increasing.

Table 4.2.2

Unlinked population and anomalous cases

	In both the estimated study population (LEED) and AES	In LEED-estimated study population but not in AES	In AES with S&W>0 but not in LEED	LEED and AES, S&W=0 in AES
	N	*N*	*N*	*N*
Average 2000 to 2004	59,530	68,740	50,910	3,210
2000	58,510	64,310	58,280	2,820
2001	61,650	63,100	52,980	4,660
2002	60,200	66,990	55,710	3,010
2003	58,880	72,570	42,180	3,080
2004	58,400	76,740	45,380	2,500
	Means	*Means*	*Means*	*Means*
Monthly employees (LEED)	12.4	7.0	0.0	6.7
Monthly employment (LEED + LBF)	13.9	8.4	1.0	8.1
Monthly employees (AES)	13.3	…	0.1	8.6
Total annual payroll ($000) (LEED)	398.3	77.8	…	188.4
Total annual payroll ($000) (AES)	393.1	…	15.7	…
Value-added ($000) (AES)	855.7	…	83.6	89.1
Persons employed	%	%	%	%
0			19.1	
1<5	72.7	50.7	80.8	79.7
5<10	13.7	27.7	0.2	12.0
10<20	6.1	14.4	0.0	4.4
20<50	3.9	6.1	0.0	2.1
50<100	1.8	0.9	0.0	0.9
100+	1.8	0.3	0.0	1.0
Industry distribution	%	%	%	%
Agriculture	35.1	5.6	31.2	30.6
Mining	0.3	0.0	0.1	0.4
Manufacturing	6.8	12.9	3.2	5.5
Electricity, gas & water	0.1	0.0	0.0	0.2
Construction	8.5	12.3	10.6	10.7
Wholesale trade	3.4	9.5	2.6	2.9
Retail trade	13.0	20.7	5.9	11.9
Accomodation, restaurants & cafes	4.7	7.4	1.2	4.2
Transport & storage	2.9	4.5	2.4	3.0
Communication	0.7	0.6	1.4	0.9
Finance & insurance	1.4	1.9	2.1	2.4
Business services	11.7	15.4	32.4	15.1
Government	0.0	0.0	0.0	0.0
Education	1.0	0.7	0.3	0.7
Health & community services	5.4	4.0	2.2	5.8
Recreational & cultural services	1.5	1.6	2.6	1.6
Personal & household services	3.8	2.9	1.7	4.1

Note: The data in this table are unweighted. The numbers in columns 1 and 2 do not exactly sum to the totals shown in column 1 of table 4.2.1 because column 1 is a point-in-time estimate of the total study population drawn from the LBF as it exists now, not the BF files actually used to draw each annual AES sample. A small number of firms that were sampled in AES are not included in our current population estimates.

Symbol:
… not applicable

Firms that report salaries and wages in AES but do not appear in LEED

A surprisingly large number of firms reported salary and wage expenditure on their AES questionnaire or IR10 return but did not have a LEED match. The number and attributes of these firms are shown in the third column of table 4.2.2. The average annual payroll reported in AES by firms in this group was $16,000, although the median salary and wage total was much lower at $2,000. Firms in agriculture and property services are over-represented.

Overall, the ‘missing’ firms do not tend to be large employers. This means that their exclusion from the linked AES-LEED sample is not as serious as it might initially seem.

It is possible that some firms classify payments made to business owners or family workers as ‘salaries and wages’ in their annual accounts and IR10 responses. These payments are not associated with a PAYE return and therefore are not recorded in LEED. This is one possible explanation for the large number of firms with AES-reported salaries and wages but no LEED employees. Another is the treatment of casual workers, who may not always have PAYE deducted from their earnings. Timing discrepancies between the two data sources is another possible explanation.

Firms with zero salaries and wages in AES but with LEED employment

A small number of firms in the ‘matched’ sample have no salary and wage expenditure reported in AES. The most likely reason is that they did simply not supply an estimate of their expenditure on salaries and wages. The profile of these firms is shown in the fourth column of table 4.2.2. The profile is reasonably similar to that of the matched sample, although there are fewer larger firms.

Comparison of common variables for the linked sample

Variables that both AES and LEED draw from the Business Frame, such as business type and industry, have a high level of correspondence in the linked AES-LEED dataset.[31] For example, the business type code differs in just 0.24 percent of cases and 6-digit industry differs in around 3 percent of cases. Legitimate differences can exist because of differences in the dates when data are drawn from the Business Frame. The LEED and LBF datasets include updates to Business Frame data that may not have been processed at the time that the AES sample was drawn.

The correspondence of two common variables, average employment and total annual salaries and wages, is analysed in more detail in table 4.2.3. It shows the distribution of the AES-LEED relative gap in responses. Specially, the ratio (AES-LEED)/(AES+LEED)/2) is calculated for each variable.

31 Because LEED is a monthly dataset, enterprise attributes like industry can change during the year. For this comparison we used the data for the first month of the financial year.

Table 4.2.3

Relative difference between the AES and LEED estimates of employee numbers and total annual salaries and wages

	Gap in mean number of employees	Gap in total annual salaries and wages	Gap in total annual salaries and wages: Postal sample	Gap in total annual salaries and wages: Tax sample	Gap in total annual salaries and wages: Firms with 10+ employees
N	117,280	201,210	35,100	166,110	29,110
Mean	-0.03	-0.05	-0.07	0.03	-0.05
p1	-1.10	-1.39	-1.38	-1.48	-1.30
p2	-0.60	-1.04	-1.05	-1.00	-0.88
p3	-0.30	-0.85	-0.87	-0.67	-0.67
p4	-0.22	-0.71	-0.73	-0.46	-0.55
p5	-0.19	-0.60	-0.64	-0.33	-0.46
p10	-0.09	-0.33	-0.35	-0.11	-0.25
p20	-0.03	-0.13	-0.15	-0.03	-0.12
p30	-0.01	-0.05	-0.07	0.00	-0.05
p40	0.00	-0.02	-0.02	0.00	-0.02
p50	0.00	0.00	0.00	0.01	0.00
p60	0.00	0.01	0.01	0.02	0.01
p70	0.00	0.03	0.03	0.04	0.03
p80	0.02	0.06	0.06	0.11	0.06
p90	0.05	0.16	0.14	0.28	0.14
p95	0.11	0.28	0.23	0.56	0.21
p96	0.14	0.34	0.28	0.66	0.25
p97	0.18	0.43	0.36	0.79	0.30
p98	0.26	0.57	0.48	0.99	0.39
p99	0.40	0.84	0.70	1.28	0.56
Within +/- 10%	85.2	62.4	61.2	68.8	64.0

Note: Each gap is calculated as (AES-LEED)/(average of AES, LEED). Estimates are weighted by sampling weights and employment.

In the first column of table 4.2.3, the 'rolling mean employees' variable in the AES dataset is compared with a comparable variable calculated using the linked LEED data. Both measures are sourced from EMS returns, but the data were drawn at different times and the LEED numbers incorporate a greater level of editing. Table 4.2.3 tabulates the relative difference, weighting the results by each firm's average employment level. The mean difference per firm was 2.8 percent of the total number of employees, and in 85 percent of cases the LEED and AES figures were within plus or minus 10 percent of each other.

The AES postal survey includes a question on annual total salary and wage expenses, and a similar variable is obtained from IR10 returns for firms in the AES tax sample. For firms with AES data on salaries and wages, it is possible to compare this estimate with the total annual salary and wage payments that are recorded in LEED. This comparison is given in the second column of table 4.2.3. Once again, the variable shown is (AES-LEED)/((AES+LEED)/2). For example, 0.25 means that the AES value was 25 percent larger than the average of the two data sources. Firms with a value of zero in the salaries and wages variable in AES are excluded from the table, because these firms are likely to be non-respondents.

The LEED and AES responses are almost identical for the mean and median firm. However, a large proportion of firms supplied a value in AES that was more than 10 percent greater or smaller than the value that we have derived from LEED. After weighting by employment size, we estimate that only around 62 percent of firms met the plus or minus 10 percent level of match quality, while the remainder have greater discrepancies.

The remaining columns of the table compare the salary and wage totals for subsets of the AES-LEED linked sample. Neither the postal nor the tax sample shows a substantially better data correspondence. Medium-sized and large firms have a slightly better AES-LEED correspondence than small firms, but the difference is not great.

There are a number of possible explanations for discrepancies between the two measures. These include:

- Errors in the allocation of employees to enterprises during the construction of LEED data. This could happen, for example, in situations where an PAYE return is supplied for a group of enterprises not a single one. There may be other types of allocation error.
- Some firms may apply different criteria to determine who is an 'employee' in their financial accounts (which tend to be the basis for AES returns) than they use in their PAYE returns.
- Some firms may classify certain types of payment as 'salaries and wages' in their financial accounts but not in PAYE returns, such as non-taxable employment-related expenses.

Firms with large AES-LEED salary and wage discrepancies

We investigated the characteristics of firms whose salary and wage expenditure in AES is very different from their expenditure on salaries and wages in LEED. We focused on the 20 percent of enterprises with the largest positive and largest negative values of (AES-LEED)/((AES-LEED)/2) for the salary and wage variable (selecting the 10 percent of records at each extreme).

A profile of these firms is given in table 4.2.4. Firms whose LEED expenditure is greater than their AES expenditure are shown in the first column. Firms whose AES expenditure is greater than their LEED expenditure are shown in the second column. Comparative data for the total sample is given in the final column. We also examined the 2-digit industry distribution of these firms, but do not show the results here for confidentiality reasons.

The group of firms with very high salary and wage totals in AES compared with LEED included a disproportionate number of businesses in farming and property services, which are more likely to be owned and operated by families. This gives some credibility to the idea that the gap may arise because working proprietors or unpaid family workers or contractors are included in salary and wage totals in annual accounts, but do not have PAYE deducted.

This group of firms has very high average labour productivity values, despite being smaller than firms in the sample as a whole. Fixed assets per person and average weekly hours are also relatively high, so it is possible that some of the higher labour productivity is genuine. The alternative explanation is that labour inputs are underestimated in the LEED-source employee and working proprietor data and this is the reason for the apparently high labour productivity.

It is less easy to understand the sources of error in the case of firms whose LEED salary and wage expenditure is greater than their AES expenditure. This group of firms is smaller than average and has somewhat lower than average estimated labour productivity (as would be the case if their LEED-matched employment levels are higher than their true employment levels). The industry distribution of this subsample is not very different from that of all firms, suggesting that there is unlikely to be any simple, over-arching explanation for the discrepancy between the LEED and AES data. However, firms in wholesale trade, property and business services are somewhat over-represented. These are industries with relatively high levels of self-employment, so perhaps the explanation partly lies in differing treatment of self-employed persons in the two data sources.[32]

Table 4.2.4

Profile of records with very divergent salary and wage data

Input or performance measure	Unit of measurement	Records where AES S&W < LEED S&W (10% lowest)	Records where AES S&W > LEED S&W (10% highest)	Total analytical sample
Value-added	Annual $(000)	337.7	724.4	766.1
Profits	Annual $(000)	111.5	84.5	207.3
Fixed assets	$(000)	301.2	938.9	724.1
Value-added per person	Annual $(000)	41.3	78.1	47.2
Value-added per hour	$ per hour	19.7	36.0	22.3
Profit per person	Annual $(000)	14.6	21.5	15.0
Fixed assets per person	$(000)	52.7	138.7	64.3
Mean employment	Average monthly total	8.6	7.0	12.8
Hours per person	Average annual total	2,109	2,220	2,108
Earnings per employee (LEED)	Annual $(000)	29.8	19.8	24.2
Employee fixed effect	Average log differential	-0.20	-0.29	-0.26
Age of employees	Average (years)	36.9	33.6	35.0
Percent workforce female	Percentage	39.9	39.4	43.0
Persons employed				
1<5	%	63.4	77.3	63.3
5<10	%	23.2	13.6	18.3
10<20	%	8.8	5.6	10.2
20<50	%	3.5	2.4	5.5
50<100	%	0.7	0.6	1.5
100+	%	0.5	0.6	1.2
Total number enterprises		20,210	20,210	202,100
Annual average sample size		4,040	4,040	40,420

Note: Sampling weights applied.

In attempt to better understand the predictors of poor match, we regressed the salary and wage match variable on a large number of firm attributes. Although many coefficients in these regressions were statistically significant, very few showed strong relationships. The total variation explained by all explanatory variables combined was also quite low (around 5 percent). This suggests that the true predictors of poor match quality are not included in our dataset.

In section 4.9, we investigate the sensitivity of our results to whether or not firms with particularly large discrepancies between their LEED and AES salary and wage data are included in the sample.

4.3 Profile of firms in the linked study sample

In the construction of the linked data, we exclude AES firms without LEED matches and a large number of firms that did not respond to AES. These steps have the potential to cause

32 The LEED employment data used in this analysis do include PAYE-administered payments made to a small group of workers who were later identified as being self-employed, on the basis of other income tax data.

sample bias (although the effect of those biases could be reduced through re-weighting). Section 4.2 considered the implications of the first step and this section focuses on the second.

Table 4.3.1

Study sample attributes

Input or performance measure	Unit of measurement	Sample for aggregate statistics (including imputed observations)	Initial analytical sample (imputed obs dropped, missing value added dropped)	Final analytical sample (retain if positive value-added and fixed assets)
Value-added per person	Annual $(000)	43.1	44.0	47.2
Value-added per hour	$ per hour	20.3	20.8	22.3
Profit per person	Annual $(000)	13.3	13.7	15.0
Profit per hour	$ per hour	6.2	6.4	7.1
Fixed assets per person	$(000)	61.7	62.1	64.3
Mean employment	Average monthly total	11.2	12.5	12.8
Mean number employees	Average monthly total	9.7	11.0	11.3
Hours per person	Average annual total	2,104	2,105	2,108
Earnings per employee (LEED)	Annual $(000)	24.0	24.2	24.2
Employee fixed effect	Average log differential		-0.3	-0.3
Age of employees	Average (years)		35.1	35.0
Percent workforce female	Percentage		43.5	43.0
Persons employed				
1<5	%	64.9	63.8	63.3
5<10	%	18.3	18.1	18.3
10<20	%	9.6	10.0	10.2
20<50	%	4.9	5.4	5.5
50<100	%	1.3	1.5	1.5
100+	%	1.0	1.2	1.2
Industrial distribution				
Agriculture	%	19.1	19.8	20.0
Mining	%	0.2	0.2	0.2
Manufacturing	%	10.5	10.7	11.0
Electricity, gas & water	%	0.0	0.0	0.0
Construction	%	11.1	10.5	10.8
Wholesale trade	%	6.7	6.9	7.0
Retail trade	%	17.1	17.0	17.2
Restaurants, cafes and hotels	%	5.7	5.2	5.1
Transport & storage	%	3.8	3.8	3.8
Communication	%	0.6	0.6	0.5
Finance & insurance	%	1.5	1.6	1.4
Business services	%	14	13.5	13.1
Education	%	0.8	0.8	0.4
Health & community services	%	4.4	4.7	4.7
Cultural services	%	1.5	1.3	1.3
Personal & household services	%	3.3	3.4	3.4
Total number enterprises		297,400	221,720	202,100
Annual average sample size		59,480	44,340	40,420

Note: Estimates were obtained using sampling weights.

Table 4.3.1 presents summary data on sample characteristics. The first column gives data on firms in the linked AES-LEED sample that met the study population firm type and employment criteria in each year. The second column profiles the sample that remained after all imputed observations, and records with zero in the value-added field, were dropped. The third column profiles the sample that remained after firms with negative value-added results and negative or zero fixed asset estimates were also dropped.

The exclusions reduce the total number of enterprises by around one third. Smaller firms were more likely to be dropped because of imputation or non-positive values of the key variables. This means that the enterprises in the analytical samples have slightly higher average values of labour productivity, profits per person, fixed assets, employment and earnings than the original sample. The industry composition of the sample is not greatly changed by the exclusions.

Table 4.3.2 reviews the contribution of study sample firms to the total economic activity of *all* private enterprises in AES. The figures in the table give the percentage of each aggregate that comes from firms in the study samples. The entire study sample, including the imputed observations (column 1) accounts for about 27 percent of firms in the AES,[33] but 84 percent of reported value-added, 62 percent of profits, and about 96 percent of salaries and wages paid. However, only about 56 percent of the total value of fixed assets (as recorded in AES) was associated with the study sample firms. The employment counts shown in the bottom rows of the table indicate that the included firms cover the vast majority of employees but only about 45 percent of working proprietors.

Table 4.3.2

Economic activity of study samples as a share of all economic activity in AES private-sector enterprises

	Sample for aggregate statistics %	Initial analytical sample %	Second analytical sample %
Number of enterprises	26.6	19.9	18.1
Value-added	83.7	70.9	71.2
Profits	62.4	53.3	55.5
Salaries and wages paid to employees (AES)	95.7	80.9	76.1
Fixed assets	56.3	47.1	44.2
Rolling mean employment (AES)	98.9	80.0	75.2
Working proprietors (LBF)	44.7	32.7	30.4
Employees (LEED)	99.7	80.6	75.8

Note: Estimates were obtained using sampling weights.

After the exclusion of firms with imputed data and other exclusions, firms remaining in the study sample continue to account for more than 70 percent of value-added and three-quarters of employees (columns 2 and 3). The samples for analysis retain a high proportion of the larger firms, which not surprisingly make a greater contribution to gross domestic output.

One option in this situation would be to re-weight the selected sample so that it better represents the intended population. There are a variety of more or less complex ways in which the reweighting could be done. Even a relatively simple post-stratification of the data using a core variable such as firm size could have a beneficial impact, reducing the effects of the record exclusions.

To summarise, the exclusion of non-matched and imputed records has an important impact on the profile of the study sample. However, the data shown here indicate that these exclusions do not necessarily undermine the value of the sample for research purposes.

33 This low proportion reflects the fact that a large number of firm in the AES sample have no employees, just working proprietors.

4.4 Labour productivity measures in aggregate

This section presents aggregate statistics on labour productivity and profitability calculated using the measures in the AES-LEED linked dataset and aggregating across enterprises. The purpose is to provide another basis for assessing the reasonableness of the measures. In this section, we retain all imputed responses in the sample to ensure it is representative of the intended population. Firms with negative or zero values of value-added or profits are also retained.[34]

Table 4.4.1

National average labour productivity and profitability estimates

	2000	2001	2002	2003	2004
Value-added per person ($000)	48.5	52.1	53.9	53.6	53.8
Profit per person ($000)	12.6	11.4	15.4	15.0	13.3
Value-added per hour ($)	24.2	26.1	27.1	26.9	27.0
Profit per hour ($)	6.3	5.7	7.8	7.5	6.7

Note: Results are given in March 2002 dollar values.

National averages of the real value-added and real profit variables are presented in table 4.4.1. The ‘per person’ measures are weighted by employment and the ‘per hour’ measures are weighted by hours. Consequently, the estimates of value-added per person (such as $48,500 per year in 2000) can be interpreted as showing the labour productivity of the firm in which the average person works. Similarly, the estimates of value-added per hour (such as $24.20 dollars per hour in 2000) can be interpreted as showing average productivity per hour, when averaged across all hours worked.

The data show a steep rise in labour productivity between 2000 and 2002, and little change between 2002 and 2004. The numbers imply that average labour productivity per person grew a total of 10.9 percent between 2000 and 2004. The growth rate in the average labour productivity per hour is 11.3 percent. These rates of growth are higher than the official labour productivity growth rate between 2000 and 2004, as recorded in the official labour productivity statistics (which is approximately 5.7 percent).

The study sample for this project, in contrast to the official productivity statistics, excludes publicly-owned organisations, non-profit organisations, and firms without employees. It also includes some industries that are not included in the ‘measured sector’ for the official productivity statistics. For these reasons we would not expect the growth rates to be the same.[35] Another key difference is that in the official productivity series, income shares are used to weight the contribution of each industry to the national aggregates. No such industry weighting is applied in table 4.4.1. Nevertheless, the big gap between the two estimated growth rates suggests that considerable care is needed before extrapolating estimates and trends from the unit record dataset to the economy as a whole.

Labour productivity and profits are increasing with firm size within the study sample, as shown in table 4.4.2.

34 There are a large number of firms with negative profit results (see section 3), and these reduce the overall means.

35 The growth rates estimated with the AES-LEED sample do not change much if industries that are not included in the ‘measured sector’ are excluded, however.

Table 4.4.2

Size variations in performance, 2000–04

Persons employed	N	Annual value-added pp ($000)	Value-added ph ($)	Annual profits pp ($000)	Profits ph ($)
		Means, weighted by persons employed			
1<5	216,480	40.4	18.7	13.2	6.1
5<10	40,800	38.6	18.8	9.4	4.6
10<20	18,110	41.0	20.3	7.9	3.9
20<50	11,670	46.3	23.0	9.2	4.6
50<100	5,200	53.8	26.7	12.5	6.1
100+	5,380	72.4	36.1	19.7	10.0
ALL	297,630	*52.6*	*25.8*	*13.6*	*6.7*

Note: Estimates are weighted by sampling weights and employment. pp = per person; ph = per hour.

Table 4.4.3 indicates that incorporated companies tend to be more productive than individual proprietorships or partnerships. On average, foreign-owned companies and joint ventures are far more productive than locally-owned firms, in this dataset.

Table 4.4.3

Ownership variations in performance, 2000–04

Business type	N	Annual value added pp ($000)	Value-added ph ($)	Annual profits pp ($000)	Profits ph ($)
		Means, weighted by persons employed			
Sole proprietor	93,790	31.9	15.7	15.0	7.4
Partnership	131,070	37.5	17.5	15.6	7.2
Limited liability company	71,830	55.3	27.3	13.0	6.5
Cooperative	100	-6.8	-3.2	4.4	2.4
Joint venture	270	100.7	46.4	-52.7	-22.1
Foreign company	570	112.1	56.9	46.5	23.1

Note: Estimates are weighted by sampling weights and total hours. pp = per person; ph = per hour.

The industry data presented in table 4.4.4 suggest that capital-intensive industries such as mining, and electricity and gas have much higher levels of value-added per person and profits per person, as one would expect. The average value of fixed assets is shown in the right-hand column as a rough indicator of capital intensity. The least productive industries in the table are education[36] and accommodation, restaurants and cafes.

36 Note that the sample covers the private sector only.

Table 4.4.4

Industry variations in performance, 2000–04

Industry	N	Annual value-added pp ($000)	Value-added ph ($)	Annual profits pp ($000)	Profits ph ($)	Fixed assets pp ($000)
		Means, weighted by hours worked				
Agriculture	95,224	47.5	19.9	15.3	6.4	159.0
Services to agriculture	5,739	31.4	16.6	6.4	3.3	37.9
Forestry	2,142	68.3	29.1	-43.6	-18.5	118.5
Fishing	1,323	67.9	24.1	15.8	5.6	109.2
Mining	852	262.2	106.6	81.1	33.1	328.8
Mining services	73	64.0	24.9	131.6	53.3	699.5
Food, beverages & tobacco	1,818	58.9	27.9	16.9	7.9	101.6
Textiles, clothing & footwear	2,423	39.6	19.0	7.7	3.7	19.0
Wood & paper	2,739	78.2	34.8	7.0	3.2	140.6
Printing & publishing	1,135	52.0	27.2	11.3	6.2	31.0
Petroleum & other mineral products	1,609	120.9	55.5	27.5	12.7	85.9
Non-metaillic mineral products	620	91.0	39.9	36.6	16.0	97.0
Metal products	2,285	63.5	28.3	11.9	5.3	54.8
Machinery	4,269	56.4	25.6	11.0	5.0	30.4
Other manufacturing	3,183	38.7	18.3	7.3	3.5	14.6
Electricity & gas	96	386.2	178.9	251.6	116.2	2498.2
Water supply	65	72.3	34.0	16.7	7.9	60.0
Construction	9,589	50.4	21.5	9.1	3.9	28.5
Construction services	15,555	39.8	18.1	7.8	3.5	15.1
Basic material wholesalers	2,081	108.1	49.3	31.1	14.2	58.4
Machinery & vehicle wholesaling	3,051	69.7	32.1	18.0	8.3	19.5
Personal & household wholesaling	4,882	63.4	31.2	18.1	8.9	20.0
Food retailing	14,400	22.6	15.0	5.7	3.8	13.7
Personal & household retailing	14,990	32.6	18.3	8.6	4.9	12.5
Vehicle retailing	9,349	36.0	17.1	5.8	2.7	16.1
Accommodation, cafes & restaurants	13,958	20.1	12.3	2.3	1.4	32.3
Road transport	6,419	57.2	23.3	6.7	2.8	49.1
Rail & water transport	245	73.6	32.0	-9.1	-4.2	133.6
Air transport	319	110.6	49.9	-32.7	-14.0	109.7
Other transport	96	49.0	22.1	1.6	0.7	67.4
Transport services	1,334	58.6	26.3	16.3	7.4	95.7
Storage	166	45.3	21.2	8.1	3.8	66.5
Communication	2,094	227.3	106.8	95.1	44.5	386.9
Finance	1,306	142.6	73.3	104.3	53.1	41.8
Insurance	301	100.9	51.1	23.5	11.2	19.5
Services to finance & insurance	2,591	71.5	36.4	19.8	10.1	12.8
Property services	10,170	75.6	36.9	21.2	10.4	168.6
Business services	24,615	46.9	24.5	10.5	5.4	13.6
Education	2,844	10.8	6.5	5.5	3.3	11.2
Health	13,339	36.6	21.5	10.4	6.1	13.8
Community services	2,603	20.4	13.6	2.8	1.9	31.4
Media	288	102.5	55.3	11.1	6.1	95.8
Libraries, museums & the arts	1,077	27.7	16.3	6.1	3.6	18.9
Sport & recreation	3,029	46.5	26.6	14.5	8.2	72.0
Personal services	10,518	27.4	15.5	5.5	3.0	15.2
Other services	823	57.4	26.0	13.9	6.3	78.3

Note: Estimates weighted by sampling weights and employment. pp = per person; ph = per hour.

4.5 Comparison with enterprise labour productivity estimates from other data sources

The estimates of value-added per hour and per person that were derived in this study can be usefully compared with estimates derived in Maré and Timmins (2006) and Law, Buckle and Hyslop (2006). Both sets of authors used GST-sourced measures of total sales and total purchases, taken from the GST or Business Activity Indicator (BAI) dataset, to construct a proxy measure of value-added. Both took their employment data from the Business Demography database, which provides a survey-based measure of the persons employed at each firm in February. Both sets of researchers also retained all types of organisation in their sample, including public sector and not-for-profit organisations and firms with working proprietors only. For greater comparability with the samples used by these researchers, we re-calculate our estimates of labour productivity using the entire AES-LEED sample (that is, relaxing the employment threshold and including all types of businesses).

Another difference is that both Maré and Timmins, and Law *et al* give averages covering the period from 1992 or 1994 through to 2003, while the AES-LEED dataset covers the years 2000 to 2004. Because there is a long-term upward trend in labour productivity, the AES-LEED estimates should be higher, all other things being the same.

Results are shown in tables 4.5.1 and 4.5.2. Table 4.5.1 compares 1-digit estimates of value-added per hour worked with those of Law *et al.* Table 4.5.2 compares 1-digit industry estimates of annual value-added per person with those of Maré and Timmins.

Table 4.5.1

Industry mean values of annual value-added per hour

	AES-LEED study population Mean ($)	AES-LEED all firms with employment Mean ($) (1)	BAI-BD Law et al Mean ($) (2)	Ratio (1)/(2)
Agriculture, fishing & forestry	20	21	22	0.93
Mining	97	97	65	1.49
Manufacturing	30	30	28	1.06
Electricity, gas & water	174	206	165	1.25
Construction	20	20	31	0.63
Wholesale trade	36	38	61	0.62
Retail trade	17	17	16	1.06
Accommodation, cafes & restaurants	12	12	16	0.77
Transport & storage	27	30	76	0.40
Communication	107	70	73	0.96
Finance & insurance	61	60	36	1.67
Property & business services	26	25	57	0.43
Government administration	…	…	99	…
Education	6	1	23	0.02
Health & community services	19	3	27	0.10
Cultural & recreation services	34	25	43	0.59
Personal & other services	17	7	24	0.27
All industries	*26*	*22*	*39*	*0.68*

Note: Estimates are weighted by sampling weights and employment.

Symbol:
… not applicable

Several points can be taken from these comparisons. First, there are large differences between the results obtained in these three studies. Among the most striking differences are those for industries with a large public sector component: government administration, education, health and community services.

If we consider industries where private businesses predominate, the results obtained by the different measurement strategies are more similar for some industries than others. The results for industries such as agriculture, manufacturing, construction, retail trade and accommodation, restaurants and cafes look as though they might be different measures of the same underlying output-to-labour-input ratio. However, the results for certain other industries, such as mining, communications, and finance and insurance, differ substantially.

Table 4.5.2

Industry mean values of value-added per person

	AES-LEED study population Mean ($000)	AES-LEED all firms with employment Mean ($000) (1)	BAI-BD Maré & Timmins Mean ($000) (2)	Ratio (1)/(2)
Agriculture, fishing & forestry	45	46	33	1.39
Mining	240	240	111	2.16
Manufacturing	64	64	51	1.25
Electricity, gas & water	378	450	275	1.63
Construction	44	44	34	1.29
Wholesale trade	78	81	112	0.72
Retail trade	30	30	30	1.00
Accommodation, cafes & restaurants	20	20	23	0.88
Transport & storage	62	69	125	0.55
Communication	214	130	45	2.90
Finance & insurance	120	118	35	3.38
Property & business services	49	47	56	0.83
Government administration	…	…	177	…
Education	11	1	24	0.04
Health & community services	31	4	35	0.13
Cultural & recreation services	58	44	57	0.77
Personal & other services	31	13	38	0.34
All industries	*53*	*43*	*56*	*0.94*

Note: Estimates are weighted by hours.

Symbol:
… not applicable

The differences are partly due to the fact that the measurement of gross output in the postal survey component of AES is more complex, and in some industries qualitatively different from the measurement of total net sales in BAI. In AES, for example:

- Exploration and development expenditure are included in gross output in the mining industry.
- Certain taxes, such as road user charges in transport and excise duties in manufacturing and wholesaling, are included in gross output.
- In financial services, the value of interest received minus interest paid is a significant component of gross output.
- In general insurance, the value of premiums received minus the value of claims paid is a significant component of gross output.
- Flows of government funding that are considered to be transfer payments (for national accounts purposes) are not included in the gross output of recipient organisations. For example, government funding to childcare providers and schools is not included in the estimated gross output of these organisations. Because government funding has a GST component, however, this source of income is likely to be included in net sales in the BAI data. This difference in the treatment of government funding affects a wide range of organisations in health, education and social services.
- In a wide range of industries, adjustments are made for changes in the value of stocks, including stocks of work in progress. Not all of these adjustments will be reflected in total sales and total purchase figures recorded in BAI.

4.6 The longitudinal structure of AES

AES was not designed to be a longitudinal sample. However, firms that are selected into the postal sample have a fairly high probability of reselection each year if they keep the same enterprise number on the Business Frame. As noted previously, sole traders and partnerships will appear in successive AES samples if they continue to submit an IR10 and this response does not fail the Statistics NZ edit checks.

An important factor that limits continuity in the sample is administrative changes in enterprise numbers. If a firm changes its legal identity or is sold to another owner it is normally assigned a new Inland Revenue number and a new enterprise number on the Business Frame, even if the fundamentals of the business are unchanged.[37]

Tables 4.6.1 and 4.6.2 provide information on the pattern of repeat responses in the study population between 2000 and 2004, using enterprise numbers to link enterprises across years. The first row of table 4.6.1 shows the 'survival rate' of all enterprises in LEED that met the selection criteria in the year ended March 2000, in each subsequent financial year. 'Survival' in this context simply means that the firm retained the same enterprise number and continued to meet the business type and employment level criteria. The survival estimates are based on pairwise comparisons (that is, a firm does not have to be present in 2001–2003 to be counted in 2004). Eighty-four percent of firms in the 2000 population survived into 2001 and continued to meet the criteria for selection. However, the survival rates decline fairly rapidly as time passes. Only 58 percent were present on the LBF and met the selection criteria in 2004.

The second row of table 4.6.1 shows the percentage of firms in the 2000 AES-LEED study sample that continued to exist on the LBF, and continued to meet the business type and employment criteria for inclusion in the study population, in each subsequent financial year. Eighty-three percent survived into the year ending March 2001 and continued to meet the criteria for selection. By 2004, only 52 percent were present in the LBF and met the selection criteria. These estimated survival rates for the 2000 AES-LEED sample decline slightly more quickly than the pattern for the population as a whole.

The third row of the table shows the proportion of the 2000 study sample firms that continued to be present in each subsequent AES-LEED study sample. This represents the majority, but not all, of those that appear to have survived on the LBF. The rate of retention in the AES sample declines more quickly than the rate of survival on the LBF, presumably because some firms move outside their original sample strata and are no longer selected.

The fourth row shows the proportion that provided a usable AES response in each year, as a percentage of the 2000 respondents. For example, 69 percent of the 2000 respondents also had a non-imputed response in 2001, but just 40 percent also had a non-imputed response in 2004. From a research viewpoint, this is the most important result, as it determines the potential for constructing a longitudinal sample.

The next rows of the table show the ratios of (a) survival in the AES sample, and (b) response in AES, to the estimated sample survival rate on the LBF. The first ratio indicates that about 90 percent of the 2000-sample firms that appear to have remained eligible on the LBF continued to be part of the AES-LEED study sample in 2004. This suggests that a relatively small proportion of firms were lost by being dropped from the sample. By contrast, non-

37 Statistics NZ automatically creates a new enterprise number on the Business Frame whenever a new business Inland Revenue number is generated. If it discovers the connection to the existing enterprise, information on the attributes and history of that existing enterprise is moved to the new enterprise number. The old enterprise number is not reinstated. This means that there is a degree of artificial 'churn' in enterprise numbers. See Fabling, 2006, for more details.

response had a much bigger impact, reducing the ratio of 'responded' to 'survived' to 57 percent in 2004.

Overall, these results suggest that the major factors reducing longitudinal retention are (a) movement out of the study population, (b) changes in enterprise numbers, preventing firms from being linked across years, and (c) non-response. We are unable to quantify the relative impact of (a) versus (b).

The second half of table 4.6.1 shows the results of doing the same analysis using the 2004 sample as the base, and linking to the 2000–03 samples. The results obtained are broadly similar.

Table 4.6.2 analyses the retention patterns of 2000 firms (a) with at least 10 employees, (b) in the postal sample, and (c) in the tax sample. Comparison with table 4.6.1 indicates that firms with at least 10 employees in 2000 had lower sample retention rates than all firms, but higher overall response rates. Fifty-one percent of respondents in 2000 were also respondents in 2004. The lower sample retention rates are due to the fact that larger firms are more likely to be surveyed in the postal component of the sample. This point is illustrated by the results shown in the second and third parts of the table. Firms in the postal component of the AES sample have a relatively lower rate of survival in the AES sample, but given this, a relatively high response rate. The opposite is true of firms in the tax component.

Table 4.6.1

Estimated survival and sample retention rates

	2000-01	2000-02	2000-03	2000-04	2000 Base sample	2001	2002	2003	2004
	%	%	%	%		Number matched observations			
Matched observations using 2000 sample as base									
Survival in BF (as viewed in LEED)	84.2	73.1	64.6	57.8	122,510	103,170	89,510	79,120	70,820
Survival of 2000 AES sample in BF popn	83.4	70.2	60.2	52.3	58,510	48,780	41,080	35,220	30,610
Survival in AES sample	78.2	64.7	54.6	46.9	58,510	45,730	37,830	31,960	27,470
Response in AES	68.9	53.8	47.1	39.5	43,730	30,120	23,530	20,600	17,280
Ratio survival in sample to BF survival					100.0	93.7	92.1	90.7	89.7
Ratio AES responses to BF survival					74.7	61.7	57.3	58.5	56.5
Matched observations using 2004 sample as base									
	2000-04	**2001-04**	**2002-04**	**2003-04**	**2000**	**2001**	**2002**	**2003**	**2004 Base sample**
	%	%	%	%	Number matched observations				
Survival in BF (as viewed in LEED)	52.5	60.3	70.0	83.0	70,820	81,220	94,380	111,840	134,770
Survival of 2000 AES sample in BF popn	57.1	63.9	73.2	84.9	33,330	37,310	42,760	49,610	58,400
Survival in AES sample	47.0	59.0	68.4	80.4	27,470	34,470	39,930	46,950	58,400
Response in AES	40.3	52.1	57.4	71.1	17,280	22,310	24,590	30,440	42,840
Ratio survival in sample to BF survival					82.4	92.4	93.4	94.6	100.0
Ratio AES responses to BF survival					51.8	59.8	57.5	61.4	73.4

Note: In this table 'response in AES' means the record was not imputed, and value-added was not set to zero (implying non-response to key variables). AES sampling weights for the base year were used to weight estimates.

Table 4.6.2

Estimated survival and sample retention rates of different components of the 2000 sample

	2000-01 %	2000-02 %	2000-03 %	2000-04 %	2000 Base sample	2001	2002	2003	2004
						Number matched observations			
Firms with employment of 10+ in 2000									
Survival in BF (as viewed in LEED)	95.3	86.8	79.5	73.5	22,400	21,350	19,450	17,800	16,460
Survival of 2000 AES sample in BF popn	95.4	85.3	76.6	70.6	8,300	7,920	7,080	6,360	5,860
Survival in AES sample	85.7	73.5	64.9	58.4	8,300	7,110	6,100	5,390	4,850
Response in AES	76.4	62.5	57.2	51.2	6,430	4,910	4,020	3,680	3,290
Ratio survival in sample to BF survival					100.0	89.8	86.2	84.7	82.8
Ratio AES responses to BF survival					77.5	62.0	56.8	57.9	56.1
Postal sample in 2000									
Survival in BF (as viewed in LEED)	92.9	84.0	77.1	71.8	8,620	8,010	7,240	6,650	6,190
Survival in AES sample	81.4	69.5	60.6	53.8	8,620	7,020	5,990	5,220	4,640
Response in AES	74.9	61.9	54.0	48.9	6,610	4,950	4,090	3,570	3,230
Ratio survival in sample to BF survival					100.0	87.6	82.7	78.5	75.0
Ratio AES responses to BF survival					76.7	61.8	56.5	53.7	52.2
Tax sample in 2000									
Survival in BF (as viewed in LEED)	81.7	67.8	57.3	49.0	49,880	40,769	33,837	28,571	24,420
Survival in AES sample	77.6	63.8	53.6	45.8	49,880	38,714	31,840	26,736	22,835
Response in AES	68.4	52.3	45.9	37.9	37,239	25,481	19,492	17,083	14,110
Ratio survival in sample to BF survival					100.0	95.0	94.1	93.6	93.5
Ratio AES responses to BF survival					74.7	62.5	57.6	59.8	57.8

Note: In this table 'response in AES' means the record was not imputed, and value-added was not set to zero (implying non-response to key variables). AES sampling weights for the base year were used to weight estimates.

One implication of these results is that researchers who wish to use AES data in longitudinal analysis face significant rates of sample attrition, reducing the size of the available sample. The impact on the profile of respondents is considered next.

Attributes of firms with longitudinal data compared with all firms

Sample attrition is typically not random and leads to sample biases. Table 4.6.3 compares firms that had AES responses in both 2000 and 2004, with the entire analytical sample of firms that responded in each year.[38] Only about 16,000 (or 39 percent) of the roughly 40,000 firms in the 2000 sample also responded in 2004. These firms represented 40 percent of the approximately 39,000 firms that responded in 2004.

The longitudinal sample firms tended to be larger in terms of employment: about 40 percent larger on average. The longitudinal sample firms were more likely to be operating in manufacturing or wholesale trade, and much less likely to be operating in the restaurants, cafes and hotels industry. On other variables, including value-added per person, profits per person, annual earnings per employee, average age and gender mix, and mean employee fixed effects, the longitudinal sample is surprisingly similar to the two cross-sectional samples. There is a reasonably good spread of firms across firm size groups and industries within the longitudinal sample.

This suggests that although a longitudinal sample of AES respondents will not be fully representative of all firms in any given year, it may not be so unrepresentative that inferences based on longitudinal analysis cannot be applied (with caution) to the entire cross-sectional population.

38 The 2000-year weights are used to weight the longitudinal sample.

Table 4.6.3

2000–2004 longitudinal sample compared with entire sample

Input or performance measure	Unit of measurement	2000-04 longitudinal sample in 2000	2000-04 longitudinal sample in 2004	Entire analytical sample in 2000	Entire analytical sample in 2004
Value-added per person	Annual $(000)	48.1	50.2	44.5	50.0
Value-added per hour	$ per hour	22.6	23.8	21.0	23.7
Profit per person	Annual $(000)	14.6	15.8	13.3	14.9
Profit per hour	$ per hour	6.9	7.6	6.3	7.2
Fixed assets per person	$(000)	58.8	65.0	62.9	71.3
Mean employment	Average monthly total	15.8	18.8	11.7	13.5
Mean number employees	Average monthly total	14.2	17.2	10.1	11.9
Hours per person	Average annual total	2,117	2,096	2,107	2,100
Earnings per employee (LEED)	Annual $(000)	23.9	25.9	22.9	25.4
Employee fixed effect	Average log differential	-0.2	-0.3	-0.3	-0.3
Age of employees	Average (years)	34.8	36.7	34.4	35.5
Percent workforce female	Percentage	42.0	42.9	43.5	43.0
Persons employed					
1<5	%	59.4	…	64.6	62.0
5<10	%	20.0	…	17.8	19.3
10<20	%	11.0	…	9.9	10.2
20<50	%	6.2	…	5.4	5.7
50<100	%	1.9	…	1.2	1.5
100+	%	1.6	…	1.0	1.3
Industrial distribution					
Agriculture	%	19.3	…	19.2	19.5
Mining	%	0.3	…	0.2	0.2
Manufacturing	%	13.3	…	11.3	10.3
Electricity, gas & water	%	0.0	…	0.0	0.0
Construction	%	10.5	…	10.8	11.5
Wholesale trade	%	9.0	…	7.1	7.3
Retail trade	%	18.1	…	18.0	17.0
Restaurants, cafes & hotels	%	3.4	…	5.2	5.3
Transport & storage	%	4.0	…	3.8	3.6
Communication	%	0.6	…	0.6	0.5
Finance & insurance	%	0.9	…	1.2	1.6
Business services	%	11.5	…	12.5	13.5
Education	%	0.3	…	0.3	0.5
Health & community services	%	5.0	…	5.0	4.5
Cultural services	%	1.0	…	1.4	1.3
Personal & household services	%	2.7	…	3.5	3.4
Number enterprises		15,760		39,950	39,450

Note: Estimates are weighted using sampling weights.

Symbol:
… not applicable

Table 4.6.4 focuses on a still more restricted sample: firms with a response to AES in every year from 2000 through to 2004. There were about 10,000 firms in this balanced longitudinal panel. In employment terms, they were on average about 50 percent larger than firms in the comparable cross-sectional sample. The average productivity and profitability of firms in this sample was higher than the cross-sectional averages, although the differences are not large. Firms in this balanced longitudinal sample were more likely to be located in industries such as manufacturing, wholesale trade and retail trade, and less likely to be operating in industries such as agriculture, and restaurants, cafes and hotels. Overall however (in terms of the measures shown in table 4.6.3 at least), the longitudinal sample is more similar to the reference population of all private-sector firms with employees than might have been expected.

Table 4.6.4

2000–2004 balanced longitudinal sample

Input or performance measure	Unit of measurement	Firms that responded every year: attributes in 2000	Firms that responded every year: annual averages 2000-2004	Second analytical sample (all years pooled)
Value-added per person	Annual $(000)	48.6	48.8	47.2
Value-added per hour	$ per hour	23.0	23.2	22.3
Profit per person	Annual $(000)	14.3	17.1	15.0
Profit per hour	$ per hour	6.8	8.1	7.1
Fixed assets per person	$ (000)	50.2	50.8	64.3
Mean employment	Average monthly total	17.2	19.3	12.8
Mean number employees	Average monthly total	15.6	17.7	11.3
Hours per person	Average annual total	2,105	2,090	2,108
Earnings per employee (LEED)	Annual $(000)	24.5	25.7	24.2
Employee fixed effect	Average log differential	-0.2	-0.2	-0.3
Age of employees	Average years	35.2	36.3	35.0
Percent workforce female	Percentage	42.4	42.8	43.0
Industrial distribution				
Agriculture	%	16.0	…	20.0
Mining	%	0.3	…	0.2
Manufacturing	%	14.7	…	11.0
Electricity, gas & water	%	0.0	…	0.0
Construction	%	10.2	…	10.8
Wholesale trade	%	9.4	…	7.0
Retail trade	%	19.1	…	17.2
Restaurants, cafes & hotels	%	3.0	…	5.1
Transport & storage	%	4.4	…	3.8
Communication	%	0.6	…	0.5
Finance & insurance	%	1.0	…	1.4
Business services	%	12.1	…	13.1
Education	%	0.2	…	0.4
Health & community services	%	5.2	…	4.7
Cultural services	%	1.1	…	1.3
Personal & household services	%	2.8	…	3.4
Total number enterprises				202,100
Annual average sample size		10,820		40,420

Note: Estimates are weighted using sampling weights.

Symbol:
… not applicable

Another issue for longitudinal research is appropriate sample weighting. Currently, only annual weights are available. In the analysis reported here, weights assigned to firms in the base year of each ‘panel’ were used. This is not ideal for longitudinal analysis.

Summarising this section, there is a high rate of attrition from the AES sample of enterprises from year to year, due to a combination of real-world business dynamics, administrative ‘churn’ on the Business Frame, the impact of AES sample selection procedures, and survey non-response. After four years, responses are likely to be available for 40–50 percent of a set of base-year respondents. Undoubtedly, non-response is non-random and will introduce bias into any longitudinal sample.

However, a comparison of the attributes of longitudinal respondents with cross-sectional respondents, on a range of input and output measures and firm attributes, suggests that longitudinal respondents are reasonably similar to cross-sectional respondents, on a range of measured attributes at least. This suggests that at least some types of exploratory research question could be usefully investigated using linked AES-LEED data and longitudinal techniques.

With further work, it might be possible to re-weight a longitudinal AES sample so that it better approximated the population of interest. The Business Frame could potentially be used to identify the number and characteristics of longitudinal enterprises. Re-weighting would reduce the impact of selection and attrition biases, but probably would not eliminate it entirely.

4.7 Correlation of AES responses and LEED measures across time

This section looks at the correlation of variables in the linked AES-LEED dataset from year to year, at firm level. A high level of variation in the data from one year to the next could be due to one of two things: high real-world variability in firm operations and performance, or high levels of measurement error. Due to the difficulty of distinguishing between the two, a high level of variation would cast doubt upon the suitability of the data for research purposes (a cautious user would be concerned about the effects of measurement error).

The correlation of employment, earnings, fixed assets, hours, labour productivity and profitability measures across time for longitudinal respondents is illustrated in table 4.7.1. The table takes all 2000 respondents in the analytical sample and shows the correlation of each variable in 2000 with the same variable in 2001 through to 2004 (using the sample of enterprises that responded in each subsequent year). Sampling weights are applied in the first half of the table. In the lower half of the table, firms are also weighted by their employment levels.

There is a fairly high level of correlation across time in the measures of labour productivity, employment, earnings, and hours per person.[39] For example, the year-to-year correlations in the 'value-added per person' variable are above 0.95 when sampling weights are applied, and mostly above 0.80 when employment weights are applied. The difference between the two sets of estimates suggests that larger firms had greater variability in their labour productivity than small firms.

The 'profit per person' and 'fixed assets per person' variables show a lower level of correlation across time. For example, when firms are weighted by employment size, the correlation in the 'profit per person' variable ranges between 0.44 and 0.74.

39 Note that the high correlation in the 'hours' variable is likely to be mainly due to the method of construction, which used industry-level data on hours worked.

Table 4.7.1

Autocorrelation in variables across time:
All 2000 respondents with responses in subsequent years

	2000	2001	2002	2003	2004
Correlations using sampling weights					
Value-added per person	1.00	0.98	0.98	0.97	0.95
Value-added per hour	1.00	0.97	0.97	0.97	0.95
Profit per person	1.00	0.89	0.92	0.97	0.93
Fixed assets per person	1.00	0.97	0.62	0.54	0.75
Mean employment	1.00	0.99	0.96	0.94	0.92
Employee fixed effect	1.00	0.92	0.86	0.82	0.79
Hours per person	1.00	0.94	0.92	0.90	0.86
Earnings per employee (LEED)	1.00	0.91	0.87	0.83	0.81
Correlations when firms are weighted by their employment					
Value-added per person	1.00	0.86	0.70	0.88	0.88
Value-added per hour	1.00	0.85	0.65	0.88	0.89
Profit per person	1.00	0.44	0.46	0.78	0.74
Fixed assets per person	1.00	0.96	0.28	0.69	0.85
Mean employment	1.00	1.00	0.98	0.96	0.92
Employee fixed effect	1.00	0.94	0.89	0.87	0.85
Hours per person	1.00	0.97	0.96	0.95	0.91
Earnings per employee (LEED)	1.00	0.96	0.93	0.92	0.91
Number of enterprises	39,950	27,210	21,390	18,720	15,760

Note: Estimates are weighted using sampling weights.

The exact correlations obtained are sensitive to the choice of a base year and to the inclusion or exclusion of 'outliers'. Overall however, these correlations suggest a fairly high level of consistency in the responses of longitudinal firms. This is somewhat reassuring, although it does not rule out the possibility of (a) persistent measurement error (whereby some firms consistently under-report or over-report a certain variable), or (b) some level of random measurement error.

4.8 Relationships among input, output and performance variables

In this section, we use the linked AES-LEED data to estimate some very simple production functions and other regression models of labour productivity, profits and earnings. The purpose is to identify whether plausible parameters are generated when the AES-LEED data is used to model relationships between inputs and outputs, or between inputs and financial performance. Because we do not know what the true parameters should be, in the New Zealand case, plausibility can only be assessed in a very general way, drawing on predictions from economic theory or past economic research (if comparable).

Economic theory and past research findings lead us to expect that:

- Capital and labour inputs should have certain quantitative relationships with output, reflecting typical production function relationships.
- Labour productivity is a function of capital inputs, so firms with higher levels of capital will tend to have higher labour productivity. In the current dataset, this should lead to a positive correlation between fixed assets per person and value-added per person, provided that the fixed assets variable is a reasonably good proxy for firms' capital stock.
- Profitability and labour productivity are likely to be positively correlated: firms that are more productive should be more profitable, all other things being equal.
- If there are economies of scale, labour productivity (and possibly profits per person) will be positively correlated with firm size.

- Labour productivity and profitability are likely to be positively correlated with the skill level of the workforce. In terms of the current dataset, this means that the average worker fixed effect and the average age of the workforce should be positively correlated with value-added and profits. We interpret them as proxy measures of the human capital of each firm's workforce.
- The average wage per employee is expected to be positively correlated with labour productivity (because if the marginal product of labour is higher in more productive firms, wages should also be higher). The average wage per employee may also be correlated with firm profitability.

Most of these hypotheses can be explored or illustrated using the linked AES-LEED dataset. However the lack of enterprise-level data on hours worked limits the scope for analysis of relationships involving wages or labour returns. We simply report estimates of the correlation between average annual earnings per employee and firm performance.

Data for all years from 2000 to 2004 are pooled for the analysis in this section. Results are weighted by the employment of each firm. These employment weights give greater weight to larger enterprises (where the typical employee is more likely to be employed).

Correlation coefficients

Table 4.8.1 presents summary statistics on the key variables used in the analysis of this section, while table 4.8.2 shows the correlations that exist between input, output and performance measures.

Table 4.8.1

Summary statistics on key variables

	N	Mean	Std dev	Min	Max
Production function aggregates					
ln value-added	202,100	7.78	13.08	-28.35	14.75
ln fixed assets	202,100	6.74	14.74	-6.91	15.19
ln employment	202,100	4.12	11.31	0.00	9.80
Other dependent variables					
ln value-added pp	202,100	3.66	4.37	-29.04	10.99
ln value-added ph	202,100	2.99	4.18	-29.78	10.41
ln profit pp	167,610	2.09	7.77	-31.60	11.20
ln earnings pp	202,100	3.24	3.15	-3.39	8.08
Other input measures					
ln fixed assets pp	202,100	2.63	8.10	-8.36	11.69
Av person fixed effect	201,810	-0.22	0.89	-2.78	5.22
Fraction female	201,810	0.44	1.43	0.00	1.00
Av age employees	201,810	34.63	33.18	1.00	102.00
ln hours pp	202,100	7.59	0.77	7.15	8.49
Persons employed					
1<5	202,100	0.18	1.88	0.0	1.00
5<10	202,100	0.12	1.59	0.0	1.00
10<20	202,100	0.12	1.59	0.0	1.00
20<50	202,100	0.13	1.67	0.0	1.00
50<100	202,100	0.08	1.33	0.0	1.00
100+	202,100	0.38	2.39	0.0	1.00

Note: pp = per person. ph = per hour. Estimates are weighted by sampling weights and employment.

Table 4.8.2

Point-in-time correlations between key variables

	Valued-added pp	Valued-added ph	Profit	Fixed assets	Employ ment	Person fixed effect	Fraction female	Av age employ ees	Hours	Earn ings
ln value-added pp	1.00	0.98	0.68	0.50	0.24	0.40	-0.24	0.15	0.34	0.68
ln value-added ph	0.98	1.00	0.67	0.47	0.30	0.42	-0.15	0.11	0.17	0.65
ln profit pp	0.68	0.67	1.00	0.42	0.09	0.28	-0.09	0.07	0.24	0.39
ln fixed assets pp	0.50	0.47	0.42	1.00	0.13	0.08	-0.24	0.04	0.34	0.29
ln employment	0.24	0.30	0.09	0.13	1.00	0.10	0.05	-0.01	-0.26	0.27
Av person fixed effect	0.40	0.42	0.28	0.08	0.10	1.00	0.05	-0.01	0.04	0.53
Fraction female	-0.24	-0.15	-0.09	-0.24	0.05	0.05	1.00	-0.02	-0.54	-0.34
Av age employees	0.15	0.11	0.07	0.04	-0.01	-0.01	-0.02	1.00	0.23	0.27
ln hours pp	0.34	0.17	0.24	0.34	-0.26	0.04	-0.54	0.23	1.00	0.36
ln earnings pp	0.68	0.65	0.39	0.29	0.27	0.53	-0.34	0.27	0.36	1.00

Note: pp = per person. ph = per hour. Estimates are weighted by sampling weights and employment.

Value-added per person is positively correlated with profits per person, fixed assets per person, total employment, the average employee fixed effect, the average age of employees, hours per person, and average earnings per employee. The correlation between value-added per person and profits per person is 0.68, and the correlation between value-added per person and fixed assets per person is 0.50. Other correlations, such as the correlation of 0.24 with total employment, are weaker. Value-added per person is negatively correlated with the fraction of the workforce that is female. All correlation coefficients are of the expected sign. It is difficult to judge what level of correlation might be expected in the absence of any measurement error.

Because the value-added per person and value-added per hour measures do not fully control for differences across firms in the quantity of labour inputs, some of the correlations shown here probably capture, in part, the effects of variations in hours worked across firms. Firms with a higher proportion of full-time workers, and firms whose employees work overtime, are likely to have higher outputs per person and higher average earnings than firms with a lower proportion of full-time workers, even if they are no more productive on an hour-for-hour basis. To some extent, variations in hours worked may also be contributing to the correlations between labour productivity on the one hand and profits, fixed assets, and worker fixed effects, on the other.

There are weaker positive associations between *profits per person* and fixed assets, employee fixed effects, and earnings.

Production function estimates

Another way of showing the cross-sectional relationship between inputs and outputs at firm level is to calculate simple production functions. A very simple Cobb-Douglas production function is:

(1) $Y = AL^{\alpha 1}K^{\alpha 2}M^{\alpha 3}$

where Y = a measure of output such as gross output, A is a constant, L is the quantity of labour, K is the quantity of capital, and M is the quantity of raw materials.

Taking logs, (1) can be estimated as a linear equation.

(2) $\ln Y_{it} = \ln A + \alpha_1 \ln L_{it} + \alpha_2 \ln K_{it} + \alpha_3 \ln M_{it} + \varepsilon_{it}$

In (2), α_1 gives the elasticity of output with respect to labour inputs, α_2 gives the elasticity of output with respect to capital inputs, and α_3 gives the elasticity of output with respect to raw material inputs. In this study, we measure output as value-added. Value-added output is net of intermediate inputs such as raw materials, and therefore M does not appear as a variable in our regressions.

Table 4.8.3 gives the results obtained when this basic linear equation is estimated, using fixed assets as a proxy for capital inputs. Detailed (3-digit) industry and year dummies are included in each regression, which means the coefficients can be interpreted as showing the link between inputs and outputs *within* industries, net of overall period effects. We include dummies for 147 3-digit industry groups.

The coefficients on labour and capital inputs indicate the percentage increase in output that is associated with a 1 percent increase in the quantity of the input. The coefficient of 0.83 on employment in the first regression, for example, implies that a 1 percent increase in employment is associated (across firms) with a 0.8 percent increase in value-added. The fixed asset coefficient of 0.23 indicates that a 1 percent increase in fixed assets is associated with a 0.2 percent increase in outputs.

Table 4.8.3

Production function estimates: initial models

	Model 1			Model 2		
ln value-added	**Coeff**	**SE**	**t**	**Coeff**	**SE**	**t**
Intercept	2.454	0.040	61.2	2.857	0.050	57.2
Ln fixed assets	0.229	0.008	29.6	0.204	0.007	28.5
Ln employment	0.832	0.012	70.6	0.848	0.011	79.9
Average person fixed effect				1.137	0.037	30.6
Average age employees				0.006	0.001	6.0
Proportion employees female				-0.384	0.028	-13.6
Time dummies	Y			Y		
Industry dummies	Y			Y		
N	201,810			201,810		
R^2	0.95			0.95		

Note: Estimates are weighted by sampling weights and employment.

Compared with the estimates reported in overseas research (see for example Mansfield, 1985, p175), the coefficient on labour inputs appears to be relatively high and the coefficient on capital inputs relatively low. This is at least partly due to the fact that the fixed asset variable is probably not a very good proxy measure of the productive capital of the firms in the sample. However, it is interesting to note that the estimated capital elasticity is higher if the model is estimated on firms in industries that are capital intensive in nature. Table 4.8.4 shows the results of the same model estimated for firms in the food and beverage manufacturing, transport, and communications industries. In these estimates, the estimated elasticity of outputs with respect to capital inputs is in the 0.29 to 0.48 range.

Table 4.8.4

Production function estimates for selected industries

	Food and beverage manufacturing			Transport			Communications		
Ln value-added	**Coeff**	**SE**	**t**	**Coeff**	**SE**	**t**	**Coeff**	**SE**	**t**
Intercept	2.165	0.172	12.6	2.668	0.205	13.0	1.768	0.123	14.4
Ln fixed assets	0.463	0.043	10.8	0.288	0.038	7.7	0.476	0.045	10.5
Ln employment	0.550	0.050	11.0	0.758	0.045	17.0	0.628	0.079	7.9
Time dummies	Y			Y			Y		
N	1,300			4,780			1,380		
R^2	0.91			0.95			0.98		

Note: Estimates are weighted by sampling weights and employment.

The AES-LEED dataset contains a number of other variables that can be used to improve estimates of the relationship between inputs and outputs. For example, the employee average fixed effect and employee mean age variables capture information on the quality of labour inputs. The estimates obtained in a whole-of-economy model when these additional variables are included are shown in the second specification in table 4.8.3. The worker fixed effect variable is measured in logs. The coefficient of 1.14 implies, therefore, that a 1 percent increase in this measure of worker 'quality' is associated with a 1.14 percent increase in output per person. The coefficient of -0.38 on the proportion of the workforce that is female implies that switching from all female to all male would be associated with a 32 percent increase in output per person. The coefficient on age indicates that a 1-year increase in the mean age of the workforce is also associated with a very small, but statistically significant, increase in output.

Table 4.8.5 shows the effect of changing the dependent variable to value-added per person and the capital inputs variable to fixed assets per person. To control for any scale-related variation in productivity parameters, the first specification in the table includes the log of employment while the second includes a set of firm size group dummies.

In the first specification, the coefficients on fixed assets, the average employee fixed effect, and the mean age and gender composition of the workforce, are the same as those previously estimated. They change slightly in the second specification. The coefficients on both the employment and the firm size variables indicate that larger firms are more productive than small firms (after variations in capital inputs and workforce quality are partially controlled for). The relationship between size and labour productivity is monotonic.

Table 4.8.5

Production function estimates: models 3 and 4

	Model 3			Model 4		
ln value-added per person	**Coeff**	**SE**	**t**	**Coeff**	**SE**	**t**
Intercept	2.857	0.050	57.2	2.844	0.052	54.2
Ln fixed assets per person	0.204	0.007	28.5	0.206	0.007	28.3
Average person fixed effect	1.137	0.037	30.6	1.092	0.036	30.1
Average age employees	0.006	0.001	6.0	0.007	0.001	7.2
Proportion employees female	-0.384	0.028	-13.6	-0.379	0.028	-13.6
ln employment	0.052	0.006	9.0			
Persons employed						
5<10				0.096	0.013	7.5
10<20				0.195	0.015	12.8
20<50				0.256	0.017	15.3
50<100				0.307	0.020	15.1
100+				0.341	0.024	14.2
Time dummies	Y			Y		
Industry dummies	Y			Y		
N	201,810			201,810		
R^2	0.58			0.59		

Note: Estimates are weighted by sampling weights and employment.

While most of these coefficients are plausible, we have no external benchmarks for assessing accuracy. We have reason to believe that capital inputs are poorly measured by the fixed assets variable, so there is at least one source of measurement error in these regressions that could be biasing the other coefficients to some degree.

Profits and earnings regression estimates

The association between inputs and profits is briefly explored by regressing profits per person on the full set of input variables. The results are shown in the left-hand section of table 4.8.6. We restrict the sample for this analysis to firms with positive profits (about 83 percent of the analytical sample) because the dependent variable is the log of profits per person.

Table 4.8.6

Profit and earnings regressions

	Ln profits per person			Ln average earnings			Ln average earnings		
	Coeff	SE	t	Coeff	SE	t	Coeff	SE	t
Intercept	1.491	0.108	13.8	2.053	0.047	43.6	1.341	0.066	20.3
Ln value-added per person							0.250	0.012	20.9
Ln fixed assets per person	0.329	0.015	22.0	0.073	0.005	13.7	0.022	0.003	7.2
Average person fixed effect	1.282	0.083	15.5	1.299	0.026	50.0	1.025	0.023	44.5
Average age employees	0.001	0.002	0.3	0.015	0.001	19.9	0.013	0.001	21.8
Proportion employees female	-0.215	0.069	-3.1	-0.613	0.020	-31.2	-0.518	0.017	-29.6
Persons employed									
5<10	-0.346	0.033	-10.3	0.143	0.009	16.3	0.119	0.008	15.4
10<20	-0.414	0.045	-9.2	0.220	0.010	22.5	0.171	0.009	19.1
20<50	-0.395	0.046	-8.6	0.266	0.011	25.0	0.202	0.010	21.0
50<100	-0.284	0.051	-5.6	0.303	0.014	21.2	0.227	0.013	17.6
100+	-0.296	0.052	-5.7	0.318	0.018	17.3	0.233	0.017	13.8
Time dummies	Y			Y			Y		
Industry dummies	Y			Y			Y		
N	167,380			201,810			201,810		
R^2	0.58			0.69			0.74		

Note: Estimates are weighted by sampling weights and employment.

The results indicate that profits are positively associated with the quantity of fixed assets and the average employee fixed effect. The employee fixed effect variable may be endogenous in this regression. (This would be the case if more profitable firms pay higher wages, and that wage premium is partly captured in the employee fixed effect variable). Given this, other estimates may be biased. The average profitability of firms in all size classes is lower than the profitability of firms with 1 to 4 persons employed. It is unclear why this should be the case.

The relationship between inputs and average earnings per person is explored in two regressions shown in table 4.8.6. Both specifications include basic input measures (fixed assets, the average person fixed effect, the average age of employees, and the proportion of the workforce that is female), and controls for firm size group, year, and 3-digit industry. The second specification also includes a measure of the firm's labour productivity.

There is a strong correlation between the average worker fixed effect and average earnings in both regressions. For example, in the first regression, a 1 percent increase in the fixed effect variable is associated with a 1.30 percent increase in average monthly earnings. Because each worker's fixed effect is calculated using information on his or her earnings at the current and previous employers, these two variables are positively correlated by construction.

The estimated labour productivity of the firm is positively associated with earnings in the second specification. Recall (from the discussion above) that this association is likely to be at least partly due to variations across firms in the full-time versus part-time composition of employment. Firms with a high proportion of part-time employees are likely to have both lower outputs per person and lower average earnings per person. Without better measures of labour inputs, we can't distinguish this compositional effect from the effects of other factors that may lead to a positive association between higher productivity and higher earnings.

There is a clear positive relationship between firm size and average earnings, indicating that larger firms either employ a greater proportion of full-time, full-month workers or tend to pay more for each hour worked. Evidence from an analysis of enterprise data collected in the Quarterly Employment Survey (QES) (Pike, 1995), suggests that both explanations probably apply. QES data show that average weekly hours tend to be longer at larger enterprises than smaller ones, and average hourly earnings also tend to be higher (ibid, pp. 32–34). Larger firms tend to have lower ratios of part-time to total employees than small and medium-size firms.

Longitudinal estimates of labour productivity and earnings

Firms are highly heterogeneous in factors such as technology, production methods, managerial skills and managerial strategy. The estimates presented above may be influenced by systematic variation in those factors. In this section, we incorporate enterprise fixed effects into the models estimated, and use the variation over time within each firm to estimate the relationships between inputs and outputs.

Table 4.8.7 re-estimates two labour productivity models, using the same sample but adding an intercept for each enterprise (which takes out the enterprise mean). The first model in table 4.8.7 can be compared with the second model in table 4.8.3. Real total value-added is the dependent variable. The second model can be compared with the results in table 4.8.5. Value-added per person is the dependent variable.

All coefficients are reduced in size when we focus on within-firm variation in labour productivity. For example, a 1 percent change in the average worker fixed effect is now associated with a 0.22 percent increase in labour productivity in the first model, and a 0.31 percent increase in the second model. These effects are far smaller than those estimated previously, suggesting that some of the effects found earlier were due to differences in other firm-specific factors. The coefficients on workforce age and gender composition are now insignificant or barely significant.[40]

Table 4.8.7

Panel estimates of labour productivity

	Ln value-added			Ln value-added per person		
	Coeff	**SE**	**t**	**Coeff**	**SE**	**t**
Intercept	4.789	0.134	35.8	3.781	0.084	45.0
Ln fixed assets	0.090	0.011	8.0			
Ln employment	0.636	0.029	22.1			
Ln fixed assets per person				0.116	0.013	9.2
Average person fixed effect	0.224	0.060	3.7	0.314	0.069	4.6
Average age employees	-0.001	0.001	-0.9	0.001	0.001	0.6
Proportion employees female	-0.080	0.033	-2.4	-0.089	0.033	-2.7
Firm size dummies				Y		
Year fixed effects	Y			Y		
Industry fixed effects	Y			Y		
Enterprise fixed effects	Y			Y		
N observations	201,810			201,810		
N enterprises	80,950			80,950		
R^2	0.984			0.854		

Note: Estimates are weighted by sampling weights and employment.

The panel estimates from the average earnings regression shown in table 4.8.6 can be compared with the cross-sectional ones shown in the right-hand columns of table 4.8.8. Once again, the coefficients are much smaller once the mean variation across enterprises is taken out. The estimates indicate, nevertheless, that average earnings are highly correlated with variations in the average person fixed effect. A 1 percent change in the average person fixed effect is associated with a 0.76 percent change in average earnings. Average earnings continue to be positively associated with the firm's estimated labour productivity, fixed assets, the average age of employees, and the proportion of employees that is male.

40 If most firms do not change their gender composition or workforce age structure much over a four-year period, it will be difficult to estimate these effects accurately in this type of specification.

Table 4.8.8

Panel estimates of average earnings

	Ln average earnings		
	Coeff	**SE**	**t**
Intercept	3.011	0.031	98.2
Ln value-added per person	0.044	0.003	13.1
Ln fixed assets per person	0.012	0.002	4.6
Average person fixed effect	0.758	0.030	25.0
Average age employees	0.008	0.000	17.1
Proportion employees female	-0.266	0.015	-17.9
Firm size dummies	Y		
Year fixed effects	Y		
Industry fixed effects	Y		
Enterprise fixed effects	Y		
N observations	201,810		
N enterprises	80,950		
R^2	0.961		

Note: Estimates are weighted by sampling weights and employment.

4.9 Sensitivity of results to the inclusion of firms with poorly-matching AES-LEED salary and wage data

Poor matching of AES records with LEED records or measurement error in LEED could lead to poor measurement of labour productivity. If measured employment is biased upwards or downwards, relative to a firm's true employment level, our estimates of labour productivity per person will be biased in the opposite direction.

This section considers the impact of records whose salary and wage expenditure in AES is very different from their expenditure on salaries and wages in LEED. We focus on the 20 percent of enterprises with the largest positive and largest negative values of (AES-LEED)/((AES-LEED)/2) for the salary and wage variable.

One way of assessing the impact of these records on estimates is to allow them to have separate intercepts and different slopes for each parameter within a regression model. This approach is illustrated in table 4.9.1. The first specification is the basic labour productivity regression shown earlier in table 4.8.5. The second specification also includes two control variables for the records with the worst salary and wage matches, and interactions between these dummies and every other right-hand-side variable. After netting out some of the variation in parameters that is coming from the firms with poor AES-LEED salary and wage matches, the overall fit of the model is higher and most of the coefficients are higher.

These results indicate that the inclusion or exclusion of firms with large discrepancies in their salaries and wages data is likely to have a statistically significant impact on estimates (although not necessarily an economically material one). It is recommended that future researchers be aware of this issue, undertake their own sensitivity analyses, and decide whether to exclude the categories of firm which are most likely to suffer from the mis-measurement of labour productivity.

Table 4.9.1

Labour productivity with and without controls for potential AES-LEED measurement error

	Ln value-added pp			Ln value-added pp		
	Basic model			Include intercepts and interactions for records with poor S&W match		
	Coeff	SE	t	Coeff	SE	t
Intercept	2.848	0.059	47.9	2.960	0.056	52.7
Ln fixed assets	0.246	0.010	25.5	0.221	0.007	29.6
Average person fixed effect	1.360	0.037	36.5	1.513	0.042	36.1
Average age employees	0.005	0.001	4.6	0.007	0.001	6.7
Proportion employees female	-0.358	0.026	-13.8	-0.388	0.025	-15.3
LEED>AES indicator				-0.688	0.138	-5.0
AES>LEED indicator				-0.036	0.116	-0.3
Time dummies (4)	Y			Y		
Industry dummies (43)	Y			Y		
Interactions for records with LEED>AES				Y		
Interactions for records with AES>LEED				Y		
N	201,810			201,810		
R^2	0.519			0.562		

Note: Estimates are weighted by sampling weights and employment.

5. Summary of the main findings

This paper has investigated the potential benefits of using unit record data from the Annual Enterprise Survey in conjunction with unit record data from the Linked Employer-Employee Database, to construct an enterprise-level dataset containing measures of labour productivity and financial performance. A trial dataset linking AES and LEED records was created and its strengths and weaknesses were assessed. The dataset included all privately-owned profit-oriented firms with employees. This section summarises the main findings of the investigation.

Strengths and weaknesses of the data sources

LEED is a rich source of employee data. It is considered to be the most accurate source of earnings data currently available. Its strengths include the monthly unit of observation and the fact that records are available for all wage and salary earners in the economy, and the enterprises that employed them. Significant limitations include the period of time covered by LEED at present (1999–2006) and the absence of a measure of hours worked per employee. The latter means that hourly wage estimates cannot be derived, either at individual or firm level. The poor measurement of the quantity of labour inputs makes it difficult to interpret results involving any variable or relationship that is likely to be influenced by variations in the quantity of labour inputs.[41] Another limitation is that working proprietor measures are still under development.

AES provides detailed measures of sales, gross output, value-added, intermediate inputs, income, expenditure, and profits. AES data are derived from responses to industry-specific postal questionnaires (the AES survey), and from the company accounts data that is provided to Inland Revenue on IR10 forms. Because of the detailed approach taken, output and value-added are likely to be measured more accurately in AES than in any other data source.

Some data limitations arise from the fact that AES uses a stratified sample and postal questionnaire to gather data from limited liability companies. The sample is stratified by industry and size of firm. The sampling fractions tend to be high for large companies, leading to the inclusion of most of these large companies, but they are much lower for small and medium-sized firms. This means that most small and medium-sized limited liability firms in

41 Improving the measurement of hours in LEED should be a priority in future LEED development.

the economy are *not* included in the AES sample. In 2004 around 10,000 firms with employees that were classified as limited liability companies, joint ventures or companies incorporated overseas, were sampled from a total population that we roughly estimate was around 90,000. At a detailed industry level, it is possible that the size of the AES sample could be too small for some types of analysis.

AES uses IR10 returns to obtain data on sole proprietorships, partnerships and businesses in agriculture. This approach leads to a fairly high level of coverage of the population, but the data quality tends to be poorer, and key variables are not measured in as much detail as in the postal AES survey.

The AES sample design means that weights must be used to derive population estimates. The current survey weights were designed for a specific purpose (to give good industry-level and national estimates of firms' financial performance and financial position), and they may not be ideal for all estimation purposes.

Other issues arise from the fact that one-third of AES records are imputed because of non-response (or because the response provided did not met Statistics NZ editing checks). This paper compared the non-imputed subsample with the total AES sample. The non-imputed subsample appears to represent a reasonably balanced cross-section of enterprises in the full sample, but researchers need to be aware of the potential for sample biases if imputed records are dropped. In particular, smaller firms are under-represented in the non-imputed sample. Consideration could be given to the development and application of modified weights, to make the non-imputed sample match the intended population more closely on key attributes such as size and industry.

Although AES gathers data on firms' assets at balance-sheet values, it was not designed to measure productive capital assets or the flow of capital services used by firms. This could be a significant limitation for some research purposes.

AES-LEED record matching

The vast majority of enterprises in AES with employees can be matched with employing enterprises in LEED using Statistics NZ's unique enterprise numbers. Although we believe these enterprise matches are largely accurate, a comparison of annual salary and wage expenditure variables (common to both AES and LEED) revealed that approximately 38 percent of enterprises had inconsistent data. Some of the possible reasons for this were discussed in this paper. One hypothesis, for example, is that payments for some types of labour inputs (such as contractors, agents on commission, and family members) may be recorded as salaries and wages in financial accounts, but do not lead to PAYE deductions, or vice versa. It is also possible that a percentage of enterprise records are incorrectly matched because of incorrect enterprise numbers or administrative changes to enterprise numbers.

Further analysis indicated that the exclusion or separate identification of firms with large discrepancies in their salaries and wages data does have a statistically significant impact on estimates, although not necessarily an economically material one. It is recommended that future researchers be aware of this issue, undertake their own sensitivity analyses, and decide whether to exclude the categories of firm which are most likely to suffer from the mis-measurement of labour productivity.

Linked sample coverage and representativeness

Because AES is a sample survey, and sampling fractions vary by size of firm and industry, weights must be used to generate valid estimates of population totals.

In this investigation, we focused on private profit-oriented businesses with employees. The linked sample (with weights) covers a large proportion of total output as measured by AES. Firms in the linked AES-LEED sample contributed approximately 84 percent of the sum of value-added in the AES survey as a whole, and about 96 percent of total salaries and wages paid. After we excluded imputed records, records with missing data for value-added, and records with negative values of value-added, the final analytical sample accounted for more than 70 percent of the AES value-added total, and 76 percent of AES-estimated total expenditure on salaries and wages.

Smaller enterprises are more likely to be dropped from the sample when imputed records and outlying values are excluded. The industry composition of the sample is not much changed by these exclusions, at two-digit level at least. Nevertheless, researchers need to be aware that any significant sample exclusions have the potential to introduce bias. They should assess the fitness of their chosen subsample for the research purpose, and consider re-weighting to reduce the impact of selection biases.

Constructed labour productivity measures using linked records

Labour productivity per person measures derived from the AES-LEED linked dataset appear to be broadly plausible. However, we currently have very little ability to validate firm-level labour productivity measures against other evidence. The official approach to productivity measurement for New Zealand as a whole uses an index number approach. There are no plans to release 'level' or dollar-value measures of output per person or per hour.

A comparison of the labour productivity estimates derived in this study with labour productivity estimates obtained by Maré and Timmins (2006) and Law, Buckle and Hyslop (2006), using Business Activity Indicator (BAI) data to construct a proxy measure of valued-added and Business Demography data to measure employment, revealed large differences in some industries. We compared 1-digit industry averages. The results obtained suggest that further work using unit record data to compare the BAI and AES measures of value-added, aimed to help users better understand the pattern of differences, should be undertaken.

Longitudinal data availability

The AES sample was designed for cross-sectional estimation purposes. However, stable enterprises in the postal sample are mainly reselected each year, many firms in the tax sample provide data repeatedly through their IR10s, and researchers can link enterprise responses across years using each firm's unique enterprise number. This means a panel sample can be constructed. A limitation on this comes from the fact that changes in enterprise numbers for legal or administrative reasons can prevent firms from being linked or lead to attrition from the sample (because no attempt is made to follow and reselect firms whose enterprise number has changed). Enterprises in the postal sample that expand or contract significantly are also less likely to be reselected in subsequent years, due to sample design features. In addition, longitudinal continuity is reduced by non-response and partial non-response.

We constructed several different longitudinal samples using the 2000 to 2004 AES data and analysed the characteristics of these samples. Sixty-nine percent of respondents in our 2000 analytical sample also had non-imputed responses in 2001 but only 40 percent had non-imputed responses in 2004. Longitudinal response continuity is higher among firms that had 10 or more employees in the base year, and among firms in the postal sample. In both cases approximately 50 percent of respondents in 2000 were also respondents in 2004. The majority of medium-sized and large firms in our main study samples had at least two responses during the 2000 to 2004 period.

Response probability is associated with firm size, so longitudinal samples invariably contain a higher proportion of larger firms. On other measured attributes, the longitudinal samples that we constructed from AES appeared to be reasonably similar to the cross-sectional samples. There were no pronounced differences in industry composition, for example. However, researchers need to be aware of the potential effects of longitudinal sample attrition and explore those effects when conducting a longitudinal analysis.

Consistency of firm responses across time and within years

In general, input and output measures are quite highly correlated across time at enterprise level, suggesting a reasonable level of consistency in responses. Our analysis of the economic relationships between inputs, outputs and performance measures also gave broadly plausible results, suggesting that the unit record AES data does capture meaningful information on these relationships and could be used in modelling firm behaviour.

6. Conclusion: The potential for using linked AES-LEED data in future research

The findings of the analysis in this paper suggest that a linked AES-LEED dataset is complete enough and of sufficiently good quality to be used in exploring a class of research problems that require longitudinal enterprise data and longitudinal measures of labour productivity.

As discussed in the introduction, there is a range of ways in which linked AES-LEED data could be used to explore the relationships between firm characteristics and behaviour, the relationships between firm behaviour and performance, or the effects of policy changes on firms. For example, samples of AES records could be used in conjunction with LEED and other data to compare the productivity of firms whose employees participate in industry training programmes with that of firms whose employees do not, or to compare the productivity of firms before and after their employees undertake industry training.

The paper has described some of the main data quality issues and potential sources of measurement error that should be considered when a study using AES or linked AES-LEED data is designed. These data quality issues include the following:

- In AES, the sampling fraction varies by ownership type, industry and size of firm, and is lowest for small and medium-sized limited liability companies. This design feature could limit the sample sizes that are available for some types of research question.
- Responses to different questions in AES are not always complete or internally consistent. Responses tend to be more complete and consistent for larger firms. Editing policies in AES, which focus on larger respondents, probably reinforce natural size variations in response quality.
- Firms do not always provide consistent responses across different administrative reporting tasks. We found differences between the AES and LEED annual salary and wage totals for a substantial minority of firms. The inclusion of firms with material data inconsistencies has the potential to bias labour productivity measures that are constructed using matched data from both sources. We recommend this issue be examined on a case-by-case basis.
- Around one third of AES records are imputed. The non-imputed subsample differs in some important respects from the total sample. Analyses that are carried out using only non-imputed records may benefit from reweighting so that the weighted sample better represents the intended population. We suggest that reweighting options be considered.
- Although longitudinal samples can be constructed from AES, they are affected by attrition. Firm births and deaths, changes of enterprise number on the Business Frame, and fairly high rates of non-response, reduce the proportion of firms that are present in the AES sample from one year to the next. Analyses that are carried out on longitudinal subsamples may benefit from reweighting so that the weighted sample better represents the intended population.

In summary, there are several known sources of measurement error, and there is a strong likelihood that any subsample of non-imputed records that is selected for analytical purposes and linked longitudinally will not perfectly represent all firms in the intended population. Researchers need to be aware of these risks and be careful in drawing inferences from a subsample of AES respondents to their intended study population. They could consider developing and using modified weights to partially compensate for the effects of sample exclusions.

Although estimates of behavioural parameters that are based on subsamples of AES firms will not be the same as those that would be obtained with a more complete sample, insights gained from an analysis of those parameters may still be generalisable, depending on the context.

The investigation in this paper was not able to answer the question of whether firm-level labour productivity measures constructed using AES value-added and LEED labour inputs data are reasonably accurate. We simply assume that because of the detailed approach taken in AES to measuring value-added, it is likely to be measured more accurately in AES than in any other data source. A comparison of our estimates with the labour productivity measures that have been derived by other researchers using a GST net sales proxy for value-added, revealed large differences in some industries. We recommend that further work comparing the BAI and AES measures of value-added at unit record level should be undertaken in order to help data users better understand the pattern of differences.

References

Abowd J M, Creecy R H and Kramarz F (2002). "Computing person and firm effects using linked longitudinal employer-employee data", *LEHD Technical Paper 2002-06*, www.lehd.census.gov.

Fabling R (2006). "What chance a longitudinal firm?" Unpublished research note, Statistics NZ, Wellington.

Law D and McLellan N (2005). "The contributions of firm entry, exit and continuation to labour productivity growth in New Zealand", *Treasury Working Paper 05/01*, Wellington.

Law D, Buckle B and Hyslop D (2006). "Towards a model of firm productivity dynamics", paper presented at the Annual Conference of the New Zealand Association of Economists, 28–30 June 2006, Wellington.

Mansfield E (1985). *Microeconomics: Theory and Applications,* fifth edition, W.W. Norton, New York.

Maré D C and Hyslop, D (2006). "Worker-firm heterogeneity and matching: An analysis using worker and firm fixed effects estimated from LEED", LEED research paper, Statistics NZ, Wellington.

Maré D C and Timmins J (2006). "Geographic concentration and firm productivity", *Motu Working Paper 06-09*, www.motu.org.nz.

Pike C (1995). "Earnings, hours and employment in small, medium and large businesses", *Labour Market Statistics 1995*, Statistics NZ, Wellington.

Appendix A: Further information on data quality

1. Non-response and imputation in AES

Imputation is carried out in AES when sampled firms do not respond. Approximately 30 percent of records are imputed because of non-response. The rates of imputation for non-response in the postal and tax components of the sample are similar. There is no item imputation in situations where the responding firm has provided the minimum set of usable data but left other fields blank, provided the edit checks are passed.

The imputation process does not impute every AES variable. Only the key output variables are imputed.

The vast majority of imputations are done using the regression method. Each variable is estimated using the unit's administrative data (GST sales, GST purchases, or FTE employment) and a scaling factor. The scaling factor is estimated using data on the relationship between the administrative variable and the variable of interest for firms in the same imputation cell. Imputation cells are defined in terms of sample component, type of questionnaire, industry, GST value, and FTE value.

A small number of records are imputed using historical or mean imputation. Historical imputation simply multiplies the unit's responses in the previous period by a scaling factor, representing the average movement of the variable for similar businesses since the previous period. Mean imputation estimates a value for a unit using the average value for a set of similar businesses.

Approximately 5 percent of non-imputed records have a 'warning' attached to them in the AES database. This indicates that the financial performance data supplied by the respondent were used but financial position variables were imputed.

Table A.1 compares imputed and non-imputed firms within the study sample of private businesses with a LEED match and employees in LEED. Small firms are much more likely to have imputed records. Consequently, there is a large difference between imputed and real respondents across all the variables shown, when the mean values are compared. At the median, the two subsamples are much more similar.

Table A.1

Comparison of imputed and non-imputed observations
2001–04 annual averages

	N	Value-added ($000)	Profits ($000)	Salaries and wages paid ($000)	Number working proprietors (LBF) N	Number FTE employees (BF) N	Total employment [(1)] N
Means							
Imputed	70,620	401	118	174	5.9	6.4	7.9
Not imputed	207,110	920	274	421	13.5	13.0	14.5
Medians							
Imputed	70,620	100	45	30	1.8	1.8	3.3
Not imputed	207,110	108	37	28	1.9	1.9	3.3

Note: Estimates were obtained using sampling weights and by pooling the 2001–04 samples.
(1) Sum of employees from LEED and working proprietors from LBF.

A comparison of industry distributions, business types and sample component (postal versus tax) did not reveal any other striking differences between imputed and non-imputed firms.

2. Negative values of value-added and profits

Table A.2 shows the distribution of the value-added and profit variables across firms in the study sample (excluding imputed records and records with zero values, which are assumed to be firms that did not provide full responses). The first and third columns show the distributions obtained using sampling weights only, while the second and fourth columns show the distributions after weighting by employment (giving larger firms greater weight).

There are a substantial number of AES records where the value-added variable is negative. Negative values are also common in the profits variable (representing 22 percent of weighted responses when sampling weights are used).

Table A.2

Distribution of value-added and profits

	Value-added pp		Profits pp	
	Sampling weights	Sampling + employment weights	Sampling weights	Sampling + employment weights
	$(000)	$(000)	$(000)	$(000)
p1	-19	-24	-57	-76
p2	-7	-3	-33	-34
p3	-2	2	-23	-21
p4	1	4	-17	-15
p5	3	6	-13	-11
p10	8	12	-5	-3
p20	15	19	0	0
p30	20	25	0	1
p40	25	31	2	3
p50	31	38	5	5
p60	38	47	9	9
p70	47	58	14	14
p80	59	74	22	22
p90	85	106	37	43
p95	114	144	57	75
p96	126	158	65	85
p97	142	178	76	105
p98	168	223	94	132
p99	226	368	135	199

Note: pp = per person. ph = per hour. Estimates were obtained using sampling weights and by pooling the 2000–04 samples.

One legitimate reason for negative value-added is that the firm was being established and had unusually low sales in the start-up period, relative to its purchases of inputs. In fact, 13 percent of firms with negative value-added indicated that they had operated for fewer than 12 months, compared with about 6 percent of the total sample.

Firms with negative values of the value-added variable are spread across a wide range of industries, as shown in table A.3. They are over-represented in finance and insurance and in education. The latter is not surprising, because the government funding received by many institutions in the education industry is not included in the calculation of value-added (see section 4.5 above).

Table A.3

Firms with negative value-added

	Firms in study sample with negative value-added %	All firms in study sample %	Negative values as percent of all firms %
Agriculture, fishing & forestry	25.4	19.8	4.7
Mining	0.2	0.2	4.6
Manufacturing	5.6	10.7	1.9
Electricity, gas & water	0.1	0.0	8.5
Construction	5.0	10.5	1.7
Wholesale trade	8.1	7.0	4.3
Retail trade	13.8	17.0	3.0
Accommodation, cafes & restaurants	6.3	5.2	4.5
Transport & storage	2.3	3.8	2.3
Communication	1.1	0.6	7.2
Finance & insurance	4.1	1.6	9.8
Property & business services	11.8	13.6	3.2
Education	9.6	0.8	45.1
Health & community services	3.0	4.7	2.4
Cultural & recreation services	2.5	1.3	7.0
Personal & other services	1.2	3.4	1.3
Total	100.0	100.0	3.7

Note: Estimates were obtained using sampling weights and by pooling the 2000–04 samples.

Firms with negative value-added tend to be smaller than other firms in AES, although the employment means indicate there are some quite large firms in this group.

Table A.4

Profile of firms with negative value-added

	Firms with negative value-added	All firms in study sample
Value-added ($000)	-591	697
Profits ($000)	-404	182
Fixed assets ($000)	1,199	707
Rolling mean employment (BF)	7.3	11.4
Mean employees (LEED)	7.2	11.0
Mean employment (LEED)	8.4	12.5
Annual average number firms	1,870	44,340

Note: Estimates were obtained using sampling weights and by pooling the 2000–04 samples.

3. Firms with part-year operations

In any one year, 5–6 percent of all AES respondents[42] indicate that they operated for fewer than 12 months. While part-year operators are spread widely across industries, they are over-represented in accommodation cafes and restaurants, communication, education services, recreation services and personal services.

An analysis of the subsequent experiences of firms that were part-year operators in 2000 indicated that 3–4 percent were also part-year operators in 2001, 2002 and 2003. This relatively low proportion suggests that the majority of firms are part-year operators in one

42 Representing 2–3 percent of enterprises after weighting by employment.

year only, presumably because they were established or closed during the year. A smaller number may be seasonal businesses.

4. Outdated working proprietor data

At the time this study was done, the working proprietor data held on the BF for continuing enterprises had not been updated since 2003, when PAYE records became the primary data source for employee numbers and the old method of updating the BF through an annual survey was discontinued. For some firms, the last update may have been even earlier, because BF maintenance procedures focused on larger, more economically significant firms. Only recently-birthed firms on the BF were likely to have post-2003 working proprietor counts.

The number of working proprietors actually included in the study sample (drawing on the outdated BF counts) was stable over the 2000 to 2004 period. By contrast, HLFS data suggest there was an increase of around 10,000–20,000 in the number of working self-employed people towards the end of this period, so the study sample numbers may be underestimating the actually number of working proprietors in 2004. Working proprietors are a particularly important source of labour inputs, relative to employees, in certain industries such as agriculture and transport, so the implications of any error in estimation are likely to be more significant for those industries.

Given the possible underestimation of working proprietors in 2004, any growth in labour productivity between 2003 and 2004 could be due, in part, to this source of error in the estimation of labour inputs.

LEED contains income-tax sourced counts of people receiving self-employment income, and is a possible alternative source of working proprietor numbers. However, at the time of writing, there were a number of outstanding issues concerning the use of these data to estimate the labour inputs of working proprietors.

5. Overestimation of labour inputs due to calendar month measurement

A simple EMS or LEED-sourced monthly employee count overstates employment at a point in time, because all part-month employees are counted in the monthly total and implicitly given the same weight as full-month employees.

Job start and end dates are imputed in LEED, and in future research these could be used to estimate firms' point-in-time employee numbers.

6. Changing balance dates

Some firms change their balance dates between years. In these situations the data for successive financial years could be overlapping or separated by a time gap. The number of firms affected is minor. No adjustment has been made for this factor in the current paper.

Appendix B: Conceptual differences between GST and other measures of value-added

The following notes are based on a Statistics NZ document at:
http://www2.stats.govt.nz
/domino/external/omni/omni.nsf/outputs/Business+Activity+(GST)+Indicator#Design

GST sales is not the same as the national accounting concept of gross output.

- Gross output is measured on an accrual basis – businesses have the option of reporting activity on a cash or accrual basis or a combination (see details below).
- Gross output measures sales plus stock change, whereas GST sales is exclusive of stock change.
- Gross output does not record sales of capital goods and services. These appear as gross fixed capital formation in expenditure on Gross Domestic Product.
- Gross output does not record sales of businesses.
- Gross output does not include subsidies, whereas GST sales includes any grants or subsidies received.
- Gross output in wholesale and retail industries records the gross margin (sales less cost of goods sold), whereas GST records gross sales.

GST purchases is not the same as the national accounting concept of intermediate consumption.

- Intermediate consumption is measured on an accrual basis – businesses have the option of reporting activity on a cash or accrual basis or a combination (see details below).
- Intermediate consumption measures purchases plus stock change, whereas GST sales is exclusive of stock change.
- Intermediate consumption does not record purchases of capital goods and services. These appear as gross fixed capital formation in expenditure on Gross Domestic Product.
- Intermediate consumption does not record purchases of businesses.

Net sales (GST sales less GST purchases) is not the same as the national accounting concept of value-added because of the conceptual differences listed above. Value-added (GDP) is defined as the value of gross output less the value of intermediate consumption.

In addition to the usual sales of goods and services the GST sales variable includes other items such as:

- Sales of second-hand assets. These are normally recorded as capital items in the balance sheet of a business' accounts.
- Sales of businesses themselves. If they are sold as a going concern the sale is zero-rated. The amount of the sale will still appear in the GST sales variable. Some very large sales which breach Statistics NZ's confidentiality rules have been removed.

In addition to purchases of goods and services used in the production process, the GST purchases variable also includes:

- Purchases of land, buildings, plant and machinery etc, referred to in the National Accounts as gross fixed capital formation, which is normally recorded in the balance sheet of a business' accounts.
- Purchases of businesses themselves. If the business is sold as a going concern the amount of the sale is not record as a GST purchase. Some very large purchases which breach Statistics NZ's confidentiality rules have been removed.