Estimating the Validity of Administrative Variables

Bakker, Bart F.M.
Statistics Netherlands / VU University Amsterdam
P.O Box 24500
2490 HA The Hague, Netherlands
bbkr@cbs.nl / b.f.m.bakker@vu.nl

Introduction

The use of administrative data in social science research and official statistics has grown. For example, in recent issues of the Dutch social science journal *Mens en Maatschappij*, over one third of the articles containing empirical research made use of registry data (Bakker 2009). At statistical bureaus the preparations for the 2011 Census are well underway, with more and more countries making use of administrative data (Valente 2010). The administrative data, also called registries, are combined by linking and applying micro-integration methods to adjust the data and make them more consistent. The outcome of these statistical processes is called a *statistical register* or simply a *register* (Bakker, 2010). In countries where register-based Censuses are produced, a growing number of official statistics is based on registers, therefore quality problems may have a huge impact on the effectiveness of the information infrastructure in society.

One problem that may occur when register data are used for research or statistics is that the concepts measured in the registry do not correspond to the desired concepts. In other words: the validity of the measurement leaves much to be desired. The measurement in registries may lack validity for various reasons: the administrative concept may differ substantially from the desired concept; the people or other entities registered may have an interest in being registered in a particular way; the registry may have a severe administrative delay; the administrative practice of the registry keeper leads to biased entries; or the way the registry keeper processes the administrative input may lead to more biased data (Bakker, 2010). The micro-integration process should correct for most of the registry errors, but it cannot prevent that some errors remain in the resulting registers.

However, although the problem of validity is often mentioned in a qualitative way, validity is seldom measured in a quantitative way. In this article, a method is presented to estimate how valid register variables are. Starting from the classical test theory (e.g. Novick, 1966) the assumption is that the measurements of validity can be distinguished from reliability by repeated measurement. The validity can then be determined by using linked survey and register data, which should measure the same concepts, and then repeat the measurement. Because it is not always possible to repeat measurements, and because it is expensive, the survey and register measurement can be conceived as two items of the same construct. This idea is elaborated in this article with an empirical example.

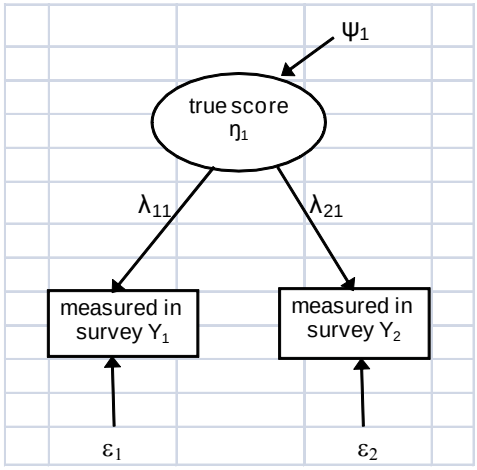
The paper starts with a short review of the literature on validity and reliability of measurement in registers and surveys. Insight into the concept of validity is enhanced by applying Linear Structural Equation Models (Jöreskog & Sörböm, 1996; Kline, 2005) with a measurement component. The construct validity of age, gender, educational attainment and wages is simultaneously determined. Section 3 describes the data of the register and survey used. Section 4 describes the results of the data analysis. Section 5 concludes on the usefulness of the method, discusses the implications for research based on administrative data, and suggests future methodological research.

Validity and reliability

In the classical test theory (Novick, 1966; Jöreskog & Sörbom, 1996; Kline, 2005), two kinds of measurement errors are distinguished: validity and reliability. According to McCall (2001) reliability refers to whether the measurement procedures assign the same value to a characteristic each time it is measured

under essentially the same circumstances. Unreliable measurement leads to random error. To estimate the reliability of a measurement instrument, it is necessary to use it twice (Figure 1). The correlation between the two measures is the estimated reliability: the test-retest-reliability. A latent variable is used for the concept that should be measured (true score η_1). In fact it is measured with Y_1 and Y_2, variables measured with errors ε_1 and ε_2. The estimated parameters λ_{11} and λ_{21} can be read as factorloadings. Their product equals the test-retest correlation. The higher the λ's, the higher the reliability and the lower the error.

Figure 1. Estimating the reliability of a survey measure

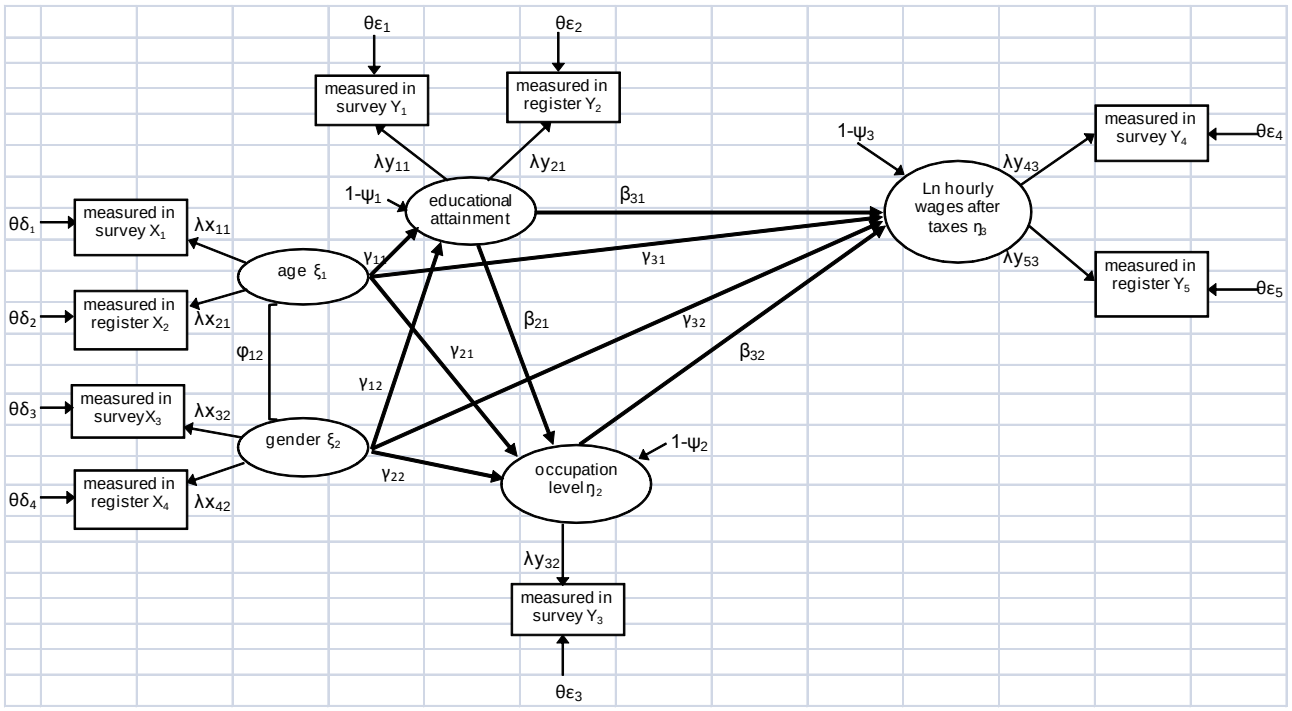


Validity refers to how accurately the values assigned in the measurement procedures reflect the actual conceptual variable measured. Invalid measurement leads to systematic error or bias in estimates (McCall, 2001). In order to estimate the validity of a measurement instrument the construct validity concept is used (McQueen & Knussen, 2002: 95-98; Singleton & Straits, 2005: 100-105). According to the logic of construct validation, the meaning of a concept is implied by the statements of its theoretical relations to other concepts. The validation process starts with the formulation of the theoretically expected relationships between variables. The more evidence that support the hypothesized relationships, the greater one's confidence that a particular measurement of the concept is valid.

In this paper, a structural equation model is used with a measurement component. Repeated measurement is available for four variables: each one from a survey and one from a register. These are the variables age, gender, educational attainment and hourly wages. However, in this case these variables are not measured with the same measurement instrument. Therefore, the correlation between the two measurements cannot be read as the test-retest reliability but as a measure for how different the measurement instruments measure the concept. Under the condition that the true scores represent the concept well, the errors ε_1 and ε_2 can be read as measures of validity.

A simple and well-known earnings function model is applied (Figure 2). The LISREL notation is used (Jöreskog & Sörbom, 1996). The target population of our research is people working more than 12 hours a week in a job. In order to have confidence in the outcomes of the model, the theoretical expectations have to be formulated. If available, the outcomes of earlier research can be used to formulate expectations on the size of the standardized effects. It is expected that age, educational attainment and occupation level have large effects on hourly wages: the expected size is between 0.30 and 0.40. Gender should have a negative effect of approximately -0.20. Furthermore, the effect of educational attainment on occupation level should be around 0.50, the effect of age and gender on occupation level should be small (approximately 0.10 and -0.10 respectively). The effect of age and gender on education level should be small and positive (around 0.10 both).

Figure 2. Model for estimating the validity of register variables age, gender, educational attainment and ln hourly wages after taxes



The data

The survey data

For the survey data, the “OSA supply panel 2004” (OSA2004) will be used. This is a household sample which is stratified according to age, gender, region and household type. The target population is people aged between 15 and 65 who do not follow daytime classes. It is a panel survey in which respondents from earlier waves are approached for a new interview. People who no longer belong to the target population are excluded. People who belong to a sampled household who did not belong to the target population in earlier waves, are now included in the sample. The interviews took place around 1 October 2004.

Age is measured by asking the birth date. From this date the age at interview date is derived. Gender is measured through the question: “What is your gender?”

Educational attainment was measured with the question: “What is the highest education program you have finished, for which you have attained a certificate?” The respondent could choose between 40 different education programmes on a show card. These programmes stem from different periods. This gives all generations the possibility to choose a suitable education programme. This information has been harmonized according to the Standard Classification of Education 2006 (SOI).

Occupation level has been measured by a rather elaborate questionnaire, asking for the job title, the most important tasks, the number of people managed, and the most important managerial tasks. The information was coded into the Netherlands Standard Classification of Occupations 1992 (Bakker, 1993). Occupation level is one of the main criteria of this classification and was derived from the occupational codes.

The wages after taxes are measured by the question "Can you tell me what your net wages are”. The interviewer first notes whether the wages are per week, per four weeks, per month or per year and then jots down the amount. Furthermore, the number of working hours is determined by the question “What are your

working hours according to your labour contract". This information always refers to the most important job in September 2004. The hourly wages were derived from the harmonized wages and working hours. This was logarithmically transformed into ln hourly wages.

The register data

The register data originates from the Social Statistical Database (SSD) of Statistics Netherlands (Bakker, 2002, 2008; Houbiers, 2004). This is a system of linked registers and surveys from 1999-2010 of which the definitive version which is adjusted by means of micro-integration is used. Micro-integration aims at improving quality by harmonizing and completing the data and adjusts the data for measurement errors. Micro-integration is executed by applying a set of decision rules. This process transforms administrative data to register data (Bakker, 2010). In this paragraph, not only the administrative sources as such, but also the micro-integration decision rules used for these four variables will be discussed. Occupation level is not measured in registries and will not be discussed in this section.

Age and *gender* data are used from the Population Register. The quality of this information is supposed to be more accurate than that from other sources. If people are not in the Population Register, Statistics Netherlands would generally use information from other sources, but here only people registered in the Population Register are used.

In the Netherlands there is no registry for *educational attainment* that covers the entire population. This is because educational registries have only recently been developed for the first time. The last time a traditional census that included information about educational attainment was held in the Netherlands was in 1971. These data are not useful for current statistics production, because the respondents can no longer be identified. Therefore, Statistics Netherlands has combined all registry data that is available, e.g. the Central Register for Enrolment in Higher Education (available since 1985), the Register of Exam Results including all pupils sitting final exams in secondary general education from 1999 onwards, the Education Number Registers for secondary general education from 2003 onwards and secondary vocational education from 2005 onwards and a few smaller registries. All these registries are recent and cover only part of the population. In particular, the population of 40 and older is not entirely covered.

To complete the population for educational attainment, the Labour Force Surveys (LFS) of 1996-2008 have been used. The Labour Force Survey is a sample survey whose target population is the population aged 15 years and older in the Netherlands, except people living in institutional households. The sample size is just under 1 percent of the population. School careers are reconstructed using the calendar method. Because the LFS is a sample survey, the resulting records should be weighted to represent the population not covered by the registries. By combining all information from current registries and surveys the educational attainment can be determined of approximately 45 percent of the population. In this paper the educational attainment measured on 30 September 2004 is used (Bakker, Linder & Van Roon, 2008). By selecting people under 50, the measured educational attainment is restricted mainly to registry entries.

Ln hourly wages from registry information starts with deriving the yearly wages after taxes of the main job. Taxes and insurance contributions are subtracted from the yearly wage before taxes registered in the fiscal administration. The wages after taxes of the main job in September 2004 is measured by taking the quotient of the yearly wages after taxes and the number of months that the main job was held. Unfortunately,

the working hours of the main job are not registered. Therefore, the working hours from the survey were used to measure hourly wages after taxes, which were logarithmically transformed to get *ln hourly wages*.

The linked dataset

In the SSD, all registers and surveys are linked to a population backbone. This is a longitudinal version of the Population Register from 1995 onwards. The most important linking variables are a personal identification number (the Netherlands' Social-Fiscal Number or Citizen Service Number), and the combination of birth date, gender, and address. In some cases surnames are used to link the data. If the data for a person changes, a new entry is made in the population backbone. All records are assigned a linking key if it can be identified in the population backbone. The OSA2004 is linked using name, birth date, gender, and address. The effectiveness is 98.9: 4730 of the 4782 respondents were assigned a linking key. The records that do not link are not very selective (Fouarge & Grim, 2007). 2873 of the 4730 linked individuals are employees with a job of more than 12 working hours a week.

The register information originates from different registries that differ in linking effectiveness. For most registries, the Social-Fiscal Number or Citizen Service Number is used for linking the data which leads to an effectiveness of over 97%. Furthermore, many entries are not linked because they do not belong to the population.

The linking key is used for linking the OSA2004 to the register data. The effectiveness is almost 100% for gender, age and ln hourly wages. However, the linking effectiveness of educational attainment is much smaller. Furthermore, the educational attainment for people aged over 50 is predominantly based on the LFS. Therefore, it cannot be used to answer the question what the validity of register information is. To restrict the impact of the number of LFS-entries persons aged under 50 are selected. After these selection processes, only 574 people could be used for the analysis. To prevent that the outcomes are biased by selection, the data are weighted to age (in ten year classes), gender and educational attainment as measured in the OSA2004-survey. If a cell contains less than three observations or if the weights were over 5.0, it was aggregated with an adjacent cell. The weights were computed with a mean of 1.0.

Table 1. Correlations between survey and register variables

	age		gender		educational attainment		occupation level	ln hourly wage after taxes	
	survey	register	survey	register	survey	register	survey	survey	register
age from survey	1.000								
age from register	.998	1.000							
gender from survey	-.070	-.072	1.000						
gender from register	-.071	-.073	.999	1.000					
educational attainment from survey	-.133	-.135	.037	.038	1.000				
educational attainment from register	-.219	-.218	.009	.010	.768	1.000			
occupation level from survey	.005	.004	-.092	-.091	.462	.529	1.000		
ln hourly wage after taxes from survey	.210	.211	-.216	-.217	.406	.427	.514	1.000	
ln hourly wage after taxes from register	.314	.313	-.188	-.188	.298	.313	.447	.823	1.000

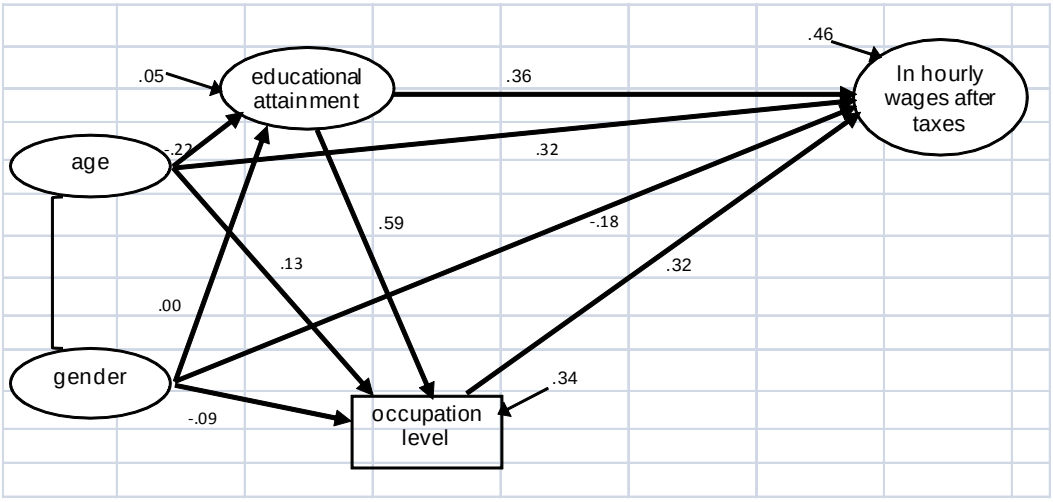
Results

Table 1 shows the correlations between the variables. As was to be expected, variables that are quite obvious measures like age and gender were measured in similar ways in the survey and the register: the

correlation is almost 1.00. However, the measurement of educational attainment and ln hourly wages are quite different.

The correlation of educational attainment from survey and register data is 0.768, while the correlation of hourly wage is only 0.823. Moreover, the educational attainment measured in the register correlates higher with occupation level and hourly wages than the version from surveys. This is true for wages measured in the survey as well as wages measured in the register. However, the differences are small except for the correlations between education and occupation (0.462 for the survey and 0.529 for the register measurement of educational attainment).

Figure 3. Evaluating the plausibility of the parameters in the model*



*The explained variance is shown for each endogenous variable (1-φ)

The validity of register variables should be demonstrated by applying a structural equation model. The complete estimated model is shown in Figure 2. This model fits the data with a χ^2 of 48 by 18 degrees of freedom. There are no important residual correlations: there is only a small residual correlation between both age variables and both wage variables and one between age and education from registers. The fit of the model did not improve much by adding parameters. Therefore the model was accepted.

In the next step the plausibility of the estimated parameters in the model is evaluated (Figure 3). If the values of the parameters are implausible then nothing could be concludes about the validity of the measured variables. However, the results corresponded with our expectations. Educational attainment (0.36), occupation level (0.32) and age (0.32) have large positive effects on wages, while gender has a moderate negative effect (-0.18). Occupation level is affected by educational attainment (0.59). Age has a moderate negative effect on educational attainment (-0.22). All other parameters are small as expected.

Table 2. Measurement errors in survey and register variables

	survey		register	
age	.00		.00	
gender	.00		.01	
educational attainment	.33	**	.11	**
occupation level	----		---	
ln hourly wages after taxes	.10	**	.24	**
Significant p<.01				

In the end, the measurement errors are evaluated (Table 2). The measurement errors of age and gender are very small and not significant. However, the errors in educational attainment and ln hourly wages are

large and significant. Educational attainment is measured with less error in the register than in the survey. The survey measurement has a significant error of 0.33, while the register measurement has an error of only 0.11. For ln hourly wages the survey is the better measurement. While the register measurement has an error of 0.24, the measurement error for the survey variable is 0.10. The differences between the size of the measurement errors is significant in both cases.

Conclusion and discussion

Despite the increased use of register data in social science research, less is known about the quality. This paper is about a method to estimate the validity of register data. With the use of the classical test theory and linear structural equations models, it is possible to quantify the construct validity. Measures from surveys can be linked to measures from registers, and under the condition that the model produces plausible results, the measurement errors can be read as a measure for the validity.

This model was applied to an earnings function, in which age, gender, education level and ln hourly wages were measured in a survey and in a register. Occupation level is also part of the model, but it is only measured in the survey. The model produces plausible results and therefore it is allowed to read the estimated measurement errors as a measure for validity. The measurement of educational attainment was better in the register than in the survey. For ln hourly wages, it is the other way around.

In this paper it is shown that the proposed method is usable in quality research of register data. Of course, one of its weaknesses is that the results depend on the knowledge of the relationships of the measured variables and other concepts. In this case there is a thorough theoretical and empirical knowledge of these relationships grounded in different disciplines like economics and sociology. However, in cases where there less knowledge it will be more difficult to apply the method. In general, it would be more difficult to apply it in a new field of research in which concepts and measurement still have to be developed.

Furthermore, it is too early to conclude anything in general about the quality of register data. This is the first attempt to estimate the validity of some register data. To come to more general conclusions, the method has to be applied to more register data. A mixed picture is expected to emerge: some variables are better measured in a particular register, others are better measured in a particular survey.

The measurement of educational attainment in the register is hybrid: most of the entries come from registers, but it is completed with entries from surveys. This shows also the inconvenience of some register data: sometimes a variable is entirely or partly missing. This urges the researcher to use survey measures to estimate the desired relationships.

REFERENCES

Bakker, B.F.M. (1993). The development of the Standard Classification of Occupations 1992, *Netherlands Official Statistics, 8* (winter 1993), 5-22.

Bakker, B.F.M. (2002). Statistics Netherlands' Approach to Social Statistics: The Social Statistical Dataset, *OECD Statistics Newsletter, 2002* (11), 4-6.

Bakker, B.F.M. (2008). De stand van het Sociaal Statistisch Bestand. *Bevolkingstrends, 56* (2), 14-18.

Bakker, B.F.M. (2009). *Trek alle registers open!* (Amsterdam: Vrije Universiteit).

Bakker, B.F.M. (2010). Micro-Integration. State of the Art (submitted)

Bakker, B.F.M., F. Linder en D. van Roon (2008). Could that be true? Methodological issues when deriving educational attainment from administrative datasources and surveys (*Shanghai: Paper prepared for the IAOS Conference on Reshaping Official Statistics, 14-16 October 2008*).

Fouarge, D. & R. Grim (2007). *Koppeling van het OSA-Arbeidsaanbodpanel aan administratieve gegevens: verslag en documentatie* (Tilburg: OSA).

Houbiers, M. (2004). Towards a Social Statistical Database and unified estimates at Statistics Netherlands, *Journal of Official Statistics, 20* (1), 55-75.

Jöreskog, K., en D. Sörbom (1996). *LISREL: 8. User's reference guide* (Chicago: Scientific Software International).

Kline, R.B. (2005). *Psychological testing: a practical approach to design and evaluation* (New York: SAGE).

McCall, R.B. (2001). *Fundamental statistics for behavioural sciences* (Belmont: Wadsworth).

McQueen, R. en C. Knussen (2002). *Research methods for social science. An introduction* (Harlow: Prentice Hall).

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3* (1), 1-18.

Singleton jr., R.A. en B.C. Straits (2005). *Approaches to social research* (Oxford / New York: Oxford University Press)

Valente, P. (2010). Main Results of the UNECE / UNSD Survey on the 2010 / 2011 Round of Censuses in the UNECE Region (Luxembourg: Eurostat)