Clustering Educational Categories in a Heterogeneous Labour Market

ROA-RM-1998/2E

Hans Heijke, Astrid Matheeuwsen, Ed Willems

This study is part of the research program 'Project Education and the Labour Market', subsidised by the Ministry of Education, Culture and Science, LDC National Centre for Career Issues, the Ministry of Social Affairs and Employment, the Ministry of Agriculture, Conservation and Fisheries and the Central Employment Board (CBA).

Research Centre for Education and the Labour Market

Faculty of Economics and Business Administration
Maastricht University

Maastricht, December 1998

ISBN 90-5321-247-7
SEC98.076/HH

Contents

Abstract

In most countries, the systems of educational classification are based on administrative criteria. For labour market analyses, however, a classification that demarcates an individual's competences obtained by the courses attended is a better alternative. In this paper we will develop an educational classification that is based on the observed substitution possibilities of workers with different educational backgrounds within similar jobs. As an additional criterion we use the recognisability of the groups distinguished. In addition, we incorporate the criterion of statistical reliability. This results in an educational classification with 113 distinct categories.

1 Introduction

In most countries, the systems of educational classification are based on administrative criteria.[1] The International Standard Classification of Education (ISCED) has such a background too. This classification – which constitutes the basis for many national classifications – distinguishes several formal levels of education and fields of study. The purpose of standard educational classifications is to reflect the formal structure of the educational system. They do not, however, indicate the real differences in competences that people have obtained during their education in order to fulfil their jobs in the labour market.

If one's objective is to give a detailed overview of all prevailing kinds of education, an administrative classification can be very useful. For most labour market research, however, the strict demarcation of educational types seems inappropriate. It ignores the segmented structure of the labour market. On the one hand, there are labour market segments with very strict educational requirements (craft markets; Doeringer and Piore, 1971), sometimes regulated by law. Examples are the requirements for doctors, lawyers, accountants *etc*. On the other hand, a large degree of educational flexibility can be observed in many segments of the labour market and there is no one-to-one relationship between education and labour market status (see *e.g.* Sheldon, 1985, and De Grip and Heijke, 1988).

The overall aim of this paper is to design an educational classification that is suitable for applied labour market research. More in particular, the classification that we intend to develop should be suitable for manpower forecasting. Parnes (1962) has already pointed out the importance of a good classification system for manpower forecasting studies. Over the years, the aim of manpower forecasting has shifted from a planning approach towards a transparency approach (see Van Eijs, 1993 and 1994). Today, manpower forecasts aim to provide useful information for (i) policy makers, who may use it to adapt the educational system, and (ii) vocational guidance for those who making educational choices. However, good classifications that take into account the actual labour market structure are still relevant. From this viewpoint De Grip, Groot, and Heijke (1991) developed an occupational classification that was based on the educational structure of the work force within occupations. The recent Standard Occupational Classification 1992 of Statistics Netherlands (CBS, 1993) also takes the educational requirements as the starting point.

It may be clear that to model the labour market developments by educational category, it is useful to use an educational classification that takes into account the actual segmentation of the labour market. It should be pointed out that we wish to focus on the competences that people acquire through education and the effects of these competences on their functioning in the labour market. To do so, we use the criterion of the actual substitution

1. With respect to occupations, Sanderson (1987) makes a distinction between functional and administrative classifications.

possibilities on the labour market, because these indicate to which extent workers with a different educational background can be employed in various occupations. These substitution possibilities implicitly indicate the overlapping skills of workers that have completed different courses. In other words, we want to take into account the substitution border lines that separate the various educational types on the labour market. This implies that the possibility of substitution is the basic criterion for developing an educational classification. In this view, substitution can be defined as the extent to which individuals with different educational backgrounds compete for the same jobs (occupations).

To develop a labour market related educational classification, we will use a clustering technique. The starting point of this cluster analysis will be the very detailed, 5-digit Dutch educational classification based on ISCED (Dutch abbreviation SOI). At this aggregation level approximately 800 educational types are distinguished. We will examine the substitution possibilities between occupational categories. The most detailed level available from the data is the 3-digit level according to ISCO 1968. At this level approximately 320 occupations are distinguished. On the basis of the above-mentioned substitution criterion, we intend to derive a classification that distinguishes approximately 100 educational categories, which seems to be a reasonable level of aggregation for both vocational guidance and most policy issues.

Besides the basic criterion of the actual demarcation lines of the labour market, we must include some additional criteria. Firstly, the aggregation level that we implement should not be too detailed, as this implies a lower statistical reliability. Hence, we take a minimum cell content of 5,000 workers for each educational category distinguished. Secondly, the educational classification should be recognisable for its users, such as policy makers, career counsellors, and individual students. Educational categories that cover various formal levels, for instance, will not be practicable for most users.

The remainder of this paper is organised as follows. In Section 2, we will discuss the primary clustering criterion, *i.e.* substitution on the labour market. Section 3 provides more insight into the data and the starting point for the clustering process. Section 4 discusses the results of the cluster analyses. In section 5, we discuss in detail the structure of the final classification, after which we will round up with the conclusion in section 6. Lastly, the entire classification and its relation with the SOI is presented in the annex.

2 Substitution on the labour market

As we have pointed out in the introduction, the labour market has a heterogeneous structure, in the sense that individuals obtain different competences and therefore have different productivity levels on the labour market. According to human capital theory (see *e.g.* Schultz, 1961, and Becker, 1962), individuals invest in 'human capital' by taking education or by obtaining experience (on-the-job training). By achieving a higher level of education

they can enhance their productivity and increase their income. More institutionalised theories, such as the labour queue theory of Thurow (1975), argue that employers select workers according to the expected training costs. Individuals with the lowest costs are placed at the head of the so-called labour queue and are therefore selected first. An individual's productivity is determined completely by the job he or she has.

The theory of job matching can be located somewhere between the two extremes of human capital theory and labour queue theory (see *e.g.* Jovanovic, 1979, and Hartog, 1992). The theory of job matching states that the productivity of individuals is neither determined completely by their jobs (labour queue), nor fully determined by their personal abilities, such as their educational background (human capital). This implies that some people (or, in our context, types of education) have a comparative advantage in the one job (occupation), whereas others have a comparative advantage in another job.

We assume that workers compete – according to their comparative advantages – for jobs with certain occupational requirements mainly on the basis of their educational background.[2] This relation between occupation and education is situated somewhere between the extremes of perfect competition on the one hand and a completely segmented one-to-one labour market on the other hand. In other words some educational types focus entirely on one or a limited number of occupations, whereas others can be used in many labour market areas.

Several studies have shown the flexibility of the various types of education by means of the Gini-Hirschman dispersion index (see *e.g*. Sheldon, 1985, De Grip and Heijke, 1988, Van der Velden and Willems, 1994, and Borghans and Heijke, 1998):

$$GH_i = \left(1 - \sum_j \left(\frac{p_{i,j}}{\sum_j p_{i,j}} \right)^2 \right) \frac{I}{I - 1} \qquad (1)$$

where:

GH_i = Gini-Hirschman dispersion index for educational category i;
$p_{i,j}$ = number of workers with educational background i in occupation j;
I = total number of educational categories distinguished.

This Gini-Hirschman index represents the realised (*ex post*) switching possibilities of

2. Workers may, however, attain additional skills by means of formal or informal training during their working career. Unfortunately we have insufficient information on such schooling or job tenures. To underpin this problem we will also restrict the analyses to youngsters – with comparable amounts of work experience – later on in this paper.

working persons with a specific educational background to other occupational classes.[3] The index is equal to 1 if and only if the workers with the educational background concerned are equally distributed across all occupations distinguished. If a type of education focuses on only one occupation, then the Gini-Hirshman index is equal to 0.

The Gini-Hirschman only indicates the occupational dispersion of the educational types. It does not provide information about other categories of workers (with different educational backgrounds) that may compete for the same occupations on the labour market. Borghans (1992) and Van der Velden and Borghans (1993) have introduced the similarity or competition index, which does provide information about the apparent substitution possibilities on the labour market. This similarity index *s* is defined as:

$$s_{i,i'} = \frac{\sum_j \left(\frac{p_{i,j}}{\sum_j p_{i,j}} \right) \left(\frac{p_{i',j}}{\sum_j p_{i',j}} \right)}{\sqrt{\sum_j \left(\frac{p_{i,j}}{\sum_j p_{i,j}} \right)^2 \sum_j \left(\frac{p_{i',j}}{\sum_j p_{i',j}} \right)^2}} \qquad (2)$$

where:

$s_{i,i'}$ = similarity index of educational category *i* with educational category *i'*.

This similarity index $s_{i,i'}$ is equal to 0 (no similarity) if the two types of education *i* and *i'* have no overlapping occupations. It is equal to 1 (perfect similarity) if and only if the occupational structure of both educational types is completely equal in the sense that the relative numbers of workers in each occupation is equal for these two educational types. If in total *I* number of educational types are distinguished, a *I x I* matrix *S* of similarity indexes can be specified. Obviously the similarity of a type of education with itself ($s_{i,i}$) is equal to 1 and the similarity index is symmetric ($s_{i,i'} = s_{i',i}$). This implies that we can distinguish *I*(*I* - 1)/2 similarity indexes.

Cluster analysis

The similarity criterion specified in equation (2) is often used in cluster analyses.[4] Clustering takes place on the basis of the highest similarity index in matrix *S*. Usually a hierarchical technique is adopted, in which in each iteration one (already clustered) educational category is combined with only one other (already clustered) category. If, for example, education *i* and education *i'* have the highest similarity $s_{i,i'}$ of all combinations, *i* and *i'*

3. Similarly, we can derive the switching possibilities to other sectors of industry or combinations of occupations and industries.

4. Other techniques are based on the (squared) distance or the correlation between two categories. See *e.g.* Lorr (1983).

together will form the new educational cluster.

After each iteration in the clustering process, we must derive the similarity index of the new cluster, say $k = i + i'$, with all other educational categories. The clustering literature distinguishes six methods: single linkage, complete linkage, average linkage, centroid clustering, medium method and minimal variance or Ward's method. All these methods state that the similarity between the new cluster k and another educational category k' ($s_{k,k'}$) is the weighted average of $s_{k',i}$, $s_{k',i'}$, and $s_{i,i'}$ (see Lorr, 1983). The weight coefficients vary over the six methods distinguished (for a discussion of the advantages and disadvantages of these methods, see De Grip, Groot and Heijke, 1987).

Although the above-mentioned methods for the calculation of new similarities between the newly formed educational clusters have obvious computational advantages, we will opt for a different – in our view less biased – technique. After every iteration in the clustering process, we recalculate similarity matrix S according to equation (2). This matrix will only be modified for the similarity indexes with the new clustered educational category. We can specify two reasons for this procedure. Firstly, the standard clustering algorithms ignore the fact that the original entities (types of education) have different sizes and thus ignore the impact on the combined similarity with other types of education. Secondly – and partly related to this – these techniques ignore the fact that the starting point of the cluster analysis is already a clustering of educational categories.

Summarising, the cluster analysis procedure can be described as follows:

1. Calculate similarity matrix S, which contains the similarity index $s_{i,i'}$ for all educational categories initially distinguished.
2. Combine the two educational categories that have the highest mutual similarity.
3. Return to step 1.

Without additional restrictions, this process will continue until only one cluster is left. Stopping rules that are generally implemented are (1) the number of categories that will eventually be distinguished, or (2) a minimum similarity required for clustering. We will stop the clustering process if the largest similarity between two educational categories is smaller than 0.5.

3 Data

The data set on which the cluster analysis described in Section 2 will be applied, is the Labour Force Survey ('Enquête Beroepsbevolking', abbreviated as EBB) 1992 and 1994[5] of Statistics Netherlands. The EBB is a continuous survey of Dutch households, focusing on the labour market situation of the labour force. Information collected includes employment status (employed, unemployed, *etc.*), educational background, sex, age and

5. Earlier EBB surveys do not include information on educational categories at the 5-digit SOI level.

for those who are employed, the sector of industry, occupation, and number of hours worked. The sample size is approximately 1%, corresponding to about 120,000 individuals.

Table 1
Number of educational categories (5-digit SOI) distinguished and average number of workers in each category by level of education, average 1992 and 1994

Level of education	Number of categories	Average number of workers per category
Primary Education	1	531,500
Lower General Secondary Education (LGSE)/ Preparatory Vocational Education (PVE)	130	10,500
Higher General Secondary Education (HGSE)/ Pre-University Education (PUE)/ Intermediate Vocational Education (IVE)	310	8,000
Higher Vocational Education (HVE)	225	4,500
University Education (UE)	177	2,500
Total (incl. rest)	844	7,000

Source: CBS/ROA

For our purpose, we subtracted from the EBB the matrix of the number of workers per educational category by occupational group on the most detailed level available. For the educational categories this implies the 5-digit SOI classification, while the occupational groups refer to 3-digit format of ISCO '68.[6] At these levels of aggregation more than 800 educational categories[7] and 320 occupational groups are distinguished. This data matrix will constitute the starting point of the cluster analysis.

To provide a better view of the data matrix used, we will first present the number of educational types distinguished per (formal) level of education in table 1. This table also gives an overview of the average number of workers in each category. Most educational categories refer to the level of Intermediate Vocational Education. At the two levels of higher education – Higher Vocational Education and University Education – however, we also distinguish many categories, with on average only 2,500 - 4,500 workers. By definition at the lowest level, Primary Education, only one type of education is distinguished.

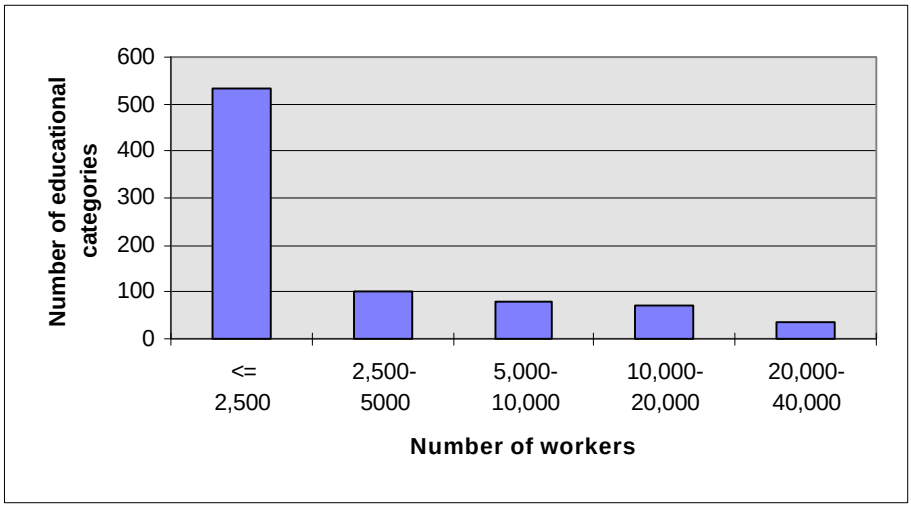
Subsequently, figure 1 presents the number educational categories by number of workers in each category. It appears that at this low aggregation level of educational specialisation, the majority of the categories represent fewer than 2,500 workers: over 500 of the 800

6. The new ISCO ’88 was not yet available during the research.

7. Excluding educational categories with no respondents in EBB.

types of education belong to this group. Within this group, the educational categories with fewer than 500 workers are overrepresented. Only 15 of the 5-digit educational categories have more than 40,000 workers.

Figure 1
Number of educational categories (5-digit SOI) by class of number of workers in each category, average 1992 and 1994



Adjustments to the data matrix of education by occupation

The large overrepresentation of very small categories in the 5-digit SOI has the disadvantage that the reliability of these figures is very low.[8] This implies that the division over the occupational groups and therefore also the values of the similarity index are more or less random for these small educational categories. To overcome these disadvantages, we have carried out a first grouping of educational categories, to ensure that groups of at least 2,500 workers are distinguished. It may be expected that the advantages of this first grouping of small categories outweigh the disadvantages of using very small categories in the cluster analysis.

This first grouping is based on the highest similarities within the same 4-digit SOI. That is to say, if an educational type at the 5-digit level has fewer than 2,500 workers, then it is clustered with the educational category within the same 4-digit category for which the mutual similarity index is highest. If necessary, this process is repeated several times. This does not yield a solution in all cases. Sometimes there is only one 5-digit category within a 4-digit category, which implies that no similar educational types can be found in this way. Furthermore, for several educational categories the similarity with the educational types within the same 4-digit category is 0 for all categories distinguished. In such cases, we made a first clustering of more or less equivalent educational categories on a more *ad hoc*

8. See CBS (1995) for the confidence intervals.

basis.

In addition to this first grouping, which was necessary to avoid educational types with only a few workers, some other categories were also taken together in advance. We first combine the various specialisations within General Secondary Education, as these subject specialisations are very difficult to measure.[9] Secondly the related graduate and post-graduate courses at university level are taken together, as this distinction was not made within the 1992 Labour Force Survey. After these two types of clustering – for categories with only a few workers and because of more institutional reasons – we distinguish fewer than 400 categories.

An additional adjustment to the data set was necessary with respect to the occupational groupings distinguished. For our purposes, a major disadvantage of the 3-digit occupational classification was that it made no distinction within the profession of teachers in secondary education. This implies that we should assume that perfect substitution can take place within this occupation, implying that teachers of English are interchangeable with teachers of Chemistry. As this is highly unrealistic, we divided the teaching professions into six main categories (General, Arts, Mathematics and Natural Sciences, Medical, Economic and Social, and Fine Arts). This was done under the assumption that teachers belong to the category in which they were educated themselves. This means that we assume that a teacher who has studied Mathematics is a teacher in the category of Mathematics and Natural Sciences.

A similar disadvantage holds for the occupational group of doctor's and dentist's assistants. It may be expected that both doctor's assistants and dentist's assistants refer to specific craft markets with hardly any substitution possibilities. We have therefore divided the category of doctor's and dentist's assistants into two occupational categories in a similar way as for the teaching professions. Lastly, we obviously excluded from the analysis any workers whose occupation was unknown.

4 Results

Free clustering of 1994 data

On the basis of the adjusted data set of EBB 1994, distinguishing 361 educational categories and 330 occupational groups, we carried out the cluster analysis according to the method described in Section 2. As a stopping criterion for clustering we specified a minimum similarity of 0.5. This means that educational categories were clustered as long

9. The category 'not further specified' is by far the largest. The three formal levels of General Secondary Education distinguished in the Dutch educational system (LGSE, HGSE and PUE) are not clustered in advance.

as there was a similar educational category (or cluster of categories) with a mutual similarity of over 0.5.

Table 2
Number of educational levels combined in one educational cluster; unrestricted cluster analysis, 1994

Educational level	Number of educational clusters
1 level	53
LGSE/PVE	5
HGSE/PUE/IVE	26
HVE	14
UE	8
2 levels	35
LGSE/PVE & HGSE/PUE/IVE	16
HGSE/PUE/IVE & HVE	10
HVE & UE	9
3 levels	11
Primary Education & LGSE/PVE & HGSE/PUE/IVE	1
LGSE/PVE & HGSE/PUE/IVE & HVE	2
HGSE/PUE/IVE & HVE & UE	8

In this analysis 103 educational clusters are distinguished. It appears that we can distinguish a large diversity of educational clusters. Some clusters consist of more than 5% or 10% of all categories distinguished, which implies over 40 5-digit categories. On the other hand, many categories consist of only one or a few categories. The picture is even more skewed if we look at the number of workers in each cluster. The largest cluster represents almost 1.8 million workers, representing approximately 30% of the total working population. In this category all levels of General Secondary Education are taken together.

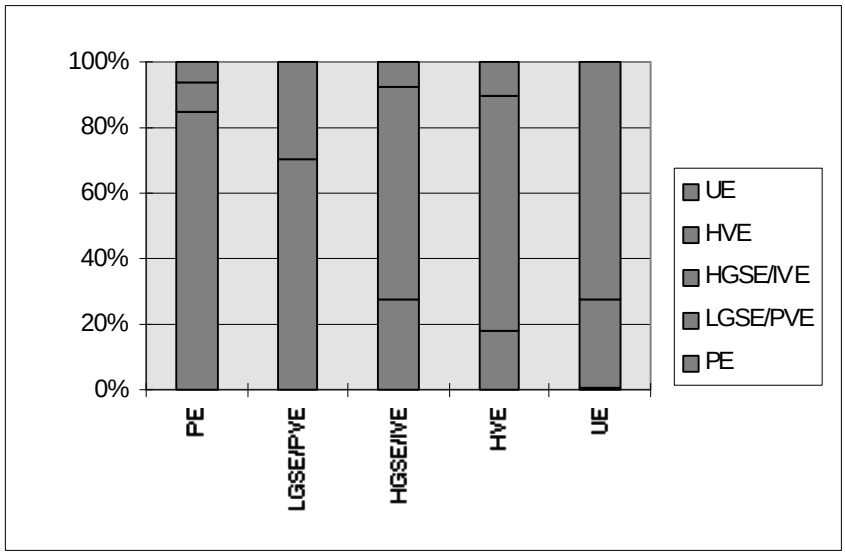
Probably the most striking result of this first cluster analysis is that much of the substitution between educational categories is not restricted to one educational level. Especially the PVE and IVE levels are frequent competitors. Table 2 shows that half of the educational clusters distinguished refer to only one formal level of education. Most of the overlapping educational levels in one cluster refer to the PVE and IVE levels on the one hand and the HVE and UE levels on the other.

In figure 2, the resulting educational clusters are classified according to the modal educational level of the cluster.[10] For each of the five levels, the distribution of the real educational level of the workers within such a category is given. As can easily be seen for the category with a modal educational level of LGSE/PVE (the second bar), about 30% of

10. For instance, if HVE has the largest share of all educational levels within a cluster, the cluster is classified at HVE level.

the labour market substitution refers to IVE-educated workers. Furthermore, competition takes place between IVE workers and workers with an educational background at HVE or even University level. Substitution between workers of HVE and University level also appears to be frequent. About 25% of the UE domain on the labour market is occupied by workers with an HVE background. Reversely, 10% of the educational clusters with HVE as the modal refers to workers with an University degree.

Figure 2
Distribution of educational levels by modal educational level of the clusters



We must, however, apply some nuances to these results. Firstly, as pointed out above, the IVE level in the SOI is very heterogeneous. Three formal educational levels are grouped in this category: 'regular' Intermediate Vocational Education (usually four years of education), 'short' Intermediate Vocational Education (SIVE; usually two years), and apprenticeships. This means that there is sometimes only a slight difference between the PVE and IVE level. Furthermore, the educational classification only registers the formal training courses. The increase of a worker's human capital due to informal on-the-job training and work experience is not reflected in the data. This implies that there may be two distinct groups within each occupation. An older group of workers with an educational background at PVE level and (unobserved) informal training and work experience, and a younger group with an IVE background. These intergenerational differences, however, cannot be seen as real intragenerational substitution possibilities between PVE- and IVE-educated workers.

Related to this phenomenon is the unobserved heterogeneity within the occupational categories distinguished. Jobs can be upgraded because of technological or organisational developments. As the data does not distinguish between the upgraded jobs and the jobs within an occupational group that are not upgraded, the observed substitution possibilities between workers of different educational levels will not indicate the real possibilities of

employers to hire workers with different educational backgrounds for the same jobs.

The above-mentioned developments may coincide: work experience or on-the-job training increases the competences of workers, as a result of which the job that they occupy can be upgraded. It is also possible, however, that the experienced worker switches to job with higher requirements. For this latter job the employer has two options: the experienced, but formally lower educated, worker or a young formally higher educated person.

Clustering by age category 1994 data

To overcome the drawback that the observed substitution possibilities are the result of heterogeneity of worker compentences within an educational category, we have restricted the analysis to the younger age groups. These workers have hardly any working experience and are therefore better comparable on the basis of their educational background as an indicator of their knowledge. In general, the results of this cluster analysis seem similar to the results of the cluster analysis described above with respect to the total working population. We now distinguish 137 educational clusters. The assumption made above (that substitution on the labour market between several educational levels is mainly the result of the fact that lower educated people with relatively much work experience and informal on-the-job training compete with younger higher educated workers for the same jobs) does not hold. Again, we see many educational clusters in which two or more educational levels are represented. Table 3 shows that one educational cluster contains all five levels distinguished. On the other hand, more than 60% of the cluster refers to only one level. Remarkably one cluster refers to the primary (LGSE/PVE) and tertiary (HVE) level. This is HVE Library and documentation together with economic specialisation within PVE.

Clustering by level of education

We believe that the results of the cluster analyses presented above cannot be used to create an educational classification. An educational cluster that refers to several formal levels of education is not practicable: policy evaluations usually distinguish between the formal educational levels and students also make their study choices within such formal levels. To overcome this drawback and with a view to the recognisability criterion outlined in the introduction, we restrict the cluster analysis in such a way that only educational categories within the same level can be combined. In this analysis we distinguish seven main categories, rather than the five formal levels distinguished in the SOI. We also make a distinction between general and vocational education. This is done because we expect that workers who have attended general education will only have had informal training for a good performance in their present jobs. As stated above, we do not observe such informal training. The seven main categories are: Primary Education, Lower General Secondary Education (LGSE), Preparatory Vocational Education (PVE), Higher General Secondary Education/Pre-University Education (HGSE/PUE), Intermediate Vocational Education (IVE),

Higher Vocational Education (HVE), and University Education (UE).

Table 3
Number of educational levels combined in one educational cluster; cluster analysis only for the younger age groups, 1994

Educational level	Number of educational clusters
1 level	85
LGSE/PVE	7
HGSE/PUE/IVE	27
HVE	30
UE	21
2 levels	
LGSE/PVE & HGSE/PUE/IVE	38
LGSE/PVE & HVE	1
HGSE/PUE/IVE & HVE	11
HVE & UE	10
3 levels	13
LGSE/PVE & HGSE/PUE/IVE & HVE	9
HGSE/PUE/IVE & HVE & UE	4
5 levels	1
Primary Education & LGSE/PVE & HGSE/PUE/IVE & HVE & UE	1

Table 4
Number of educational clusters distinguished per cluster analysis

Data year	Age group	Number of educational clusters
1994	15-64	163
1994	15-29	197
1992	15-64	177
1992 and 1994	15-64	155

With this demarcation of main educational categories, we carried out four cluster analyses. The first two analyses are similar to the two presented above, *i.e.* a cluster analysis for the total working population (15-64 years old) and a cluster analysis restricted to the

youngsters (15-29 years old), both based on 1994 data. In addition, analyses were carried out on the data set of EBB 1992 and a combination of EBB 1992 and EBB 1994, the latter for reasons of statistical reliability (larger average cell contents) and stability of the results. Table 4 gives an overview of the number of educational clusters resulting from the various analyses.

Differences in the results

Comparing the results of the various cluster analyses presented above is not a straightforward process. We therefore created an index that indicates the degree of relationship between two classifications. If slightly adjusted, the similarity index presented in Section 2 is also suitable for this purpose. Let the index $s_{c,c'}$ be defined as:

$$s_{c,c'} = \frac{\sum_i \left(\frac{p_{c,i}}{\sum_i p_{c,i}} \right) \left(\frac{p_{c',i}}{\sum_i p_{c',i}} \right)}{\sqrt{\sum_i \left(\frac{p_{c,i}}{\sum_i p_{c,i}} \right)^2 \sum_i \left(\frac{p_{c',i}}{\sum_i p_{c',i}} \right)^2}} \qquad (3)$$

where:

$s_{c,c'}$ = similarity index of educational cluster *c* with educational cluster *c'*;
$p_{c,i}$ = number of workers with educational background *i* in educational cluster *c*.

The index $s_{c,c'}$ is equal to 1 if both cluster *c* and cluster *c'* consist of exactly the same educational categories, whereas it is equal to 0 if clusters *c* and *c'* have no educational categories in common. With the aid of the indexes, we can test to what extent the results of two cluster analyses differ. Let *C* and *C'* denote the results of two cluster analyses. The total similarity of the two classifications *C* and *C'* can now be defined as:

$$TS_{C,C'} = \sum_{c \in C} \max_{c' \in C'} s_{c,c'} \qquad (4)$$

or in words: the index *TS* is the sum of the highest similarities per educational cluster. If *C* and *C'* distinguish the same educational clusters, then:

$$\max s_{c,c'} = 1 \quad \forall \quad c \in C,\ c' \in C' \qquad (5)$$

which implies that the maximum value of $TS_{C,C'}$ is equal to the number of clusters distinguished in classification *C (NUMC)*. Therefore we define:

$$\alpha_{C,C'} = \frac{TS_{C,C'}}{NUMC} \qquad (6)$$

The index α will have a value between 0 and 1. It is equal to 1 if and only if the two classifications *C* and *C'* are perfectly identical. Table 5 presents the values of α for the comparison of the results of the three alternative cluster analyses with the results for the cluster analysis for 1994 data, restricted by educational main category. From these results we may conclude that the results of the various cluster analyses are quite stable. For reasons of reliability of the data set we opt for the analyses on the basis of the combined 1992 and 1994 data set, total work force. Obviously, the average cell content in this set is highest.

Table 5
Results of stability cluster analyses: 1994 data total workforce compared with three alternatives

Data year	Alternative age groups	α
1994	15-29	0.84
1992	15-64	0.90
1992 and 1994	15-64	0.92

Some adjustments for an educational classification

After carrying out the cluster analysis, the next step is to build the new educational classification on the basis of the results. In doing so, some additional criteria have to be taken into account. The first concerns the reliability of the data. As a minimum constraint we use 5,000 workers in each category. This implies that clusters with fewer than 5,000 workers must be combined. As before, the similarity between educational clusters is taken as the criterion. After this combining of clusters, 120 types of education (of the 155 that resulted from the cluster analysis) are distinguished.

Subsequently a plausibility test is carried out. This is essentially based on the criterion of recognisability as formulated in the introduction section. This leads to several important adjustments to the initially obtained results. Firstly, in some cases the criterion of similarity or overlapping occupations leads to undesirable results. This is especially the case with occupations that require more general skills and knowledge, such as drivers, sales assistants, telephonists, and receptionists. These occupations recruit from various educational categories, which implies that these educational categories are incorrectly combined.

Secondly, we have divided a few very large clusters that resulted from the analysis. In this case a division along the lines of main branches of study has been made: agriculture, technical, economic, medical/care *etc.* The clearest example refers to the educational categories of Commerce and Community Care at PVE level. Due to the occupation of shop

assistant which they have in common, many types of education are combined. In order to demarcate the main branches more clearly, we have divided this cluster into two new categories IVE Administration and IVE Community care.

The third reason for adjusting the results of the cluster analysis relates to the fact that these results – or better the source data which were classified according to SOI – no longer correspond to the present formal educational structure. A good example of such an adjustment is the agricultural sector at PVE level. The cluster analyses distinguish three categories: 'arable farming and cattle-raising', 'market garden and flower culture' and 'layout and maintenance'. Although this seems to be a very adequate division, it has been rendered out of date by the changes in the educational system. At the moment, only one specialisation within PVE agriculture exists. We therefore make no distinction in the classification. Similar adjustments were made for the agricultural educational categories at IVE level and the social types of education within HVE.

Lastly, some adjustments were made because of the fact that no manpower forecasts could be made for the resulting clusters. This is for instance the case if a category contains a great deal of non-initial courses (*e.g.* firm-internal training courses). In such cases, we decided to separate these from the regular types of education. After these changes made to the outcomes of the cluster analysis, the new educational classification consists of 113 categories (including several categories 'not earlier classified'). The annex gives a complete overview.

5 Evaluation

In this paper, we have built an educational classification that is better suitable for labour market analyses. The background of this classification is the economic theory on labour market segmentation and the role of an individual's competences on the substitution possibilities in the labour market. This results in a classification that gives a clearer picture of the real demarcations between educational categories on the labour market. As a result, the labour market developments within the categories will be more homogeneous.

At the same time, the analyses carried out also shed more light on the substitution processes that take place on the labour market. In particular workers with an educational background at IVE level frequently compete with workers who only have a PVE education, although it may be expected that IVE workers will upgrade their jobs compared to PVE educated workers.

Unfortunately, the new classification does not solve all shortcomings of the older more administrative one. This is mainly due to the fact that the old classification has to be (at a very detailed level) the input for the new one. The most important drawback is that no distinction could be made between the various educational levels within the extensive IVE

level. As the SOI classification does not distinguish between short IVE, long IVE and apprenticeships, these are also combined within one educational level in the classification presented here.

To build this educational classification, we could only make use of data from 1992. However, the substitution possibilities between educational categories may change over time. Also, current possibilities may be the result of labour market tensions that have caused adjustment processes to take place. We therefore suggest that a similar analysis will be carried out in a few years to investigate the stability of the results and probably to update the classification.

References

Becker, G.S. (1962), 'Investment in Human Capital: A Theoretical Analysis', *Journal of Political Economy*, pp. 9-49.

Borghans, L. (1992), *A Histo-Topographic Map of Dutch University Studies*, ROA-W-1992/5E, Maastricht.

Borghans, L., H. Heijke (1998), 'Flexibility and Structure of the Dutch Labour Market' in: H. Heijke, L. Borghans (eds.), *Towards a Transparant Labour Market for Educational Decisions*, Ashgate, Aldershot/Brookfield USA/Singapore/Sydney, pp. 119-150.

CBS (1993), *Standaard Beroepenclassificatie 1992* ('Standard Occupational Classification 1992'), CBS, Voorburg/Heerlen.

Doeringer, P.B., M.J. Piore (1971), *Internal Labor Markets and Manpower Analysis*, Heath, Lexington.

Eijs, P. Van (1993), *The Manpower Requirements Approach: Background and Methodology*, ROA-RM-1993/3E, Maastricht.

Eijs, P. Van (1994), *Manpower Forecasting in the Western World: the Current State of the Art*, ROA-RM-1994/1E, Maastricht.

Grip, A. De, J.A.M. Heijke (1988), *Labour Market Indicators: an Inventory*, ROA-W-1988/1E, Maastricht.

Grip, A. De, L.F.M. Groot, J.A.M. Heijke (1987), *Clustering Occupational Classes by Educational Structure*, ROA-W-1987/2E, Maastricht.

Grip, A. De, L.F.M. Groot, J.A.M. Heijke (1991), 'Defining occupational groupings by educational structure', *Environment and Planning A*, 23, pp. 59-85.

Hartog, J. (1992), *Capabilities, Allocation and Earnings*, Kluwer Academic Publishers, Boston, Dordrecht, London.

Heijke, J.A.M. (ed.) (1994), *Forecasting the Labour Market by Occupation and Education*, Kluwer

Academic Publishers, Boston, Dordrecht, London.

Jovanovic, B. (1979), 'Job Matching and the Theory of Turnover', *Journal of Political Economy*, vol. 87, no. 5, pp. 972-990.

Lorr, M. (1983), *Cluster Analysis for Social Scientists*, Jossey-Bass, San Francisco, CA.

Parnes, H.S. (1962), *Forecasting Educational Needs for Economic and Social Development*, OECD, Paris.

Sanderson, J. (1987), 'Defining Functional Occupational Groupings', *Environment and Planning A*, vol. 19, pp. 1199-1220.

Schultz, T.W. (1961), 'Investment in Human Capital', *American Economic Review*, vol. 51, pp. 1-17.

Sheldon, G. (1985), *Die berufliche und geographische Flexibilität*, Ph.D. dissertation, Institut für Arbeitsmarkt- und Berufsforschung des Bundesanstalt für Arbeit, Nürnberg.

Thurow, L.C. (1975), *Generating Inequality*, Macmillan, New York.

Velden R.K.W. Van Der, and L. Borghans (1993), *Competition on the Labour Market. An Analysis of the Position of Types of Training*, ROA-RM-1993/5E, Maastricht.

Velden R. Van Der, E. Willems (1994), 'The Market Position of Occupations and Types of Education in the Netherlands', in: H. Heijke (ed.), *Forecasting the Labour Market by Occupation and Education*, Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 137-163.

Annex Educational classification

No.	Educational type	SOI-code	Number of workers av. 1992, 1994
	Primary Education		
001	Primary Education	00000-20199	500,000
	PVE, LGSE		
	LGSE		
002	LGSE	30100-30199	409,000
	PVE Agriculture		
003	PVE Agriculture	30637, 32100-32999	76,000
	PVE Technical trades		
004	PVE Construction trades	33610-33615,33617-33629	135,000
005	PVE Utilities installation	33616	9,000
006	PVE Mechanical trades	33631,33633, 33638	131,000
007	PVE Micro-mechanical trades	33635	5,000
008	PVE Automobile trades	33641	49,000
009	PVE Electrical trades	33650-33659	60,000
010	PVE Printing trades	33660-33669	7,000
011	PVE Bakery trades	33676	9,000
012	PVE Hotel, catering and food trades	33677, 34430, 38140, 38340	35,000
013	PVE Transport	34100-34399	36,000
014	PVE Other technical trades	33100-33199, 33642-33649, 33670-33672, 33674-33675 33678-33684, 33686-33699	37,000
	PVE Economics		
015	PVE Administration, commerce and textile	33673, 33685, 36110-36699	131,000
	PVE Community care		
016	PVE Community care	30667, 30671, 35100-35499, 37100-37199, 38100-38139, 38180-38199, 38330	184,000
	PVE Public order and security		
017	PVE Security	39100-39399	10,000
	PVE Other		
018	PVE Other	30690, 31100-31199, 31610,38430,38600-38699, 39800-39999	13,000

No.	Educational type	SOI-code	Number of workers av. 1992, 1994
	IVE, HGSE		
	HGSE		
019	HGSE	40100-40199	306,000
	IVE Agriculture and the natural environment		
020	IVE Arable farming and cattle breeding	42110-42129, 42150-42169, 42188-42199,42220-42999	112,000
021	IVE Environment	42130, 42175, 42183	22,000
	IVE Technology		
022	IVE Laboratory	43100-43199,45210-45219	19,000
023	IVE Construction	43611-43615,43618-43619 43811-43815,43818-43819	135,000
024	IVE Civil engineering	43623-43625,43823-43825	21,000
025	IVE Utilities installation	43616, 43816	25,000
026	IVE Mechanical engineering and technology	42210, 43631, 43633, 43638, 43643, 43647-43648, 43833, 43838,43843, 43847-43848, 44231, 44250-44299	140,000
027	IVE Micro-mechanical technology	43635, 43835, 45280	15,000
028	IVE Automobile technology	43641, 43649, 43841	62,000
029	IVE Aviation technology	43645	6,000
030	IVE Operational technology	43646	9,000
031	IVE Electrical technology	43650-432659,43850-43859	143,000
032	IVE Printing technology	43660-43669, 43867-43869	39,000
033	IVE Process technologies	43674, 43675, 43678, 43679, 43874, 43875, 43878, 43879	20,000
034	IVE Bakery technology	43676, 43876	16,000
035	IVE Food technology/meat processing	43677, 43877	25,000
036	IVE Transport	40657, 44100-44199, 44210, 44238-44241, 44248, 44300-44999	46,000
037	IVE Other technology	43671-43673, 43682, 43688-43699, 43888-43999	20,000

No.	Educational type	SOI-code	Number of workers av. 1992, 1994
	IVE Services and health care		
038	IVE Doctor's, dentist's and veterinary assistant	45118, 45138, 45160	18,000
039	IVE Pharmacist's assistant	45220	16,000
040	IVE Nursing	45120, 45380	97,000
041	IVE Medical technology	45230-45239	6,000
042	IVE Social and cultural	47100-47199	55,000
043	IVE Community care	41610, 45430, 45480, 48110-48129 48180-48199,48320	199,000
044	IVE Hairdressers	48130, 48330	57,000
045	IVE Hotel and catering	48140, 48340	70,000
046	IVE Physical therapy	40667, 45140-45159, 45161-45179, 45181-45199, 45240-45249, 45410, 45900	18,000
	IVE Economics		
047	IVE Administration	46110, 46133, 46138, 46141, 46148, 46180-46199, 46600-46629, 46690	269,000
048	IVE Commerce	43617, 43685, 43817, 43885 46143, 46150-46159, 46280	281,000
049	IVE Secretarial	45180, 45310, 46131	93,000
050	IVE Tourism and recreation	46142	15,000
051	IVE Business administration	46120-46129, 46220	30,000
052	IVE Computer technology	46135	28,000
053	IVE Monetary, banking and taxation	46145, 46630	36,000
054	IVE Insurance	46146	22,000
	IVE Public order and security		
055	IVE Public order and security	49100-49799	80,000
	IVE Other		
056	IVE Other	40617-40656, 40658, 40670-40699,41110-41199, 46240, 48400-48699, 49800-49999	13,000

No.	Educational type	SOI-code	Number of workers av. 1992, 1994
	HVE		
	HVE Education, Interpreter and translator		
057	HVE Primary school teacher training	40611, 50611-50617, 50690-50699, 60611-60617, 60692	144,000
058	HVE Language teacher training	50620-50625, 50628, 50629, 60620-60625, 60628, 60629, 70621-70625, 70628, 70629	27,000
059	HVE Science teacher training	50630-50659, 60630-60659, 70630-70635	33,000
060	HVE Economics and social science teacher training	50626, 50627, 50660-50666, 57181, 60626, 60627, 60660-60666 70626, 70627, 70660-70666	31,000
061	HVE Physical education teacher training	50667, 60667	15,000
062	HVE Medical and care teacher training	50670-50679, 60670-60679	14,000
063	HVE Arts teacher training	50680-50689, 60680-60689, 70680-70689	32,000
064	HVE Interpreter and translator	51100-51199	8,000
	HVE Agriculture		
065	HVE Arable farming and cattle breeding	52110, 52158	6,000
066	HVE Environmental science and food technology	52128-52130, 52171-52299, 52900, 53677, 53678	13,000
	HVE Technology		
067	HVE Laboratory	53140, 53149, 53160, 55210, 55220	38,000
068	HVE Construction	53610-53619, 53810-53819	19,000
069	HVE Civil engineering	53620-53629, 53820-53829	16,000
070	HVE Mechanical engineering	53631-53649, 53830-53849	31,000
071	HVE Electronic technology	53110-53139, 53150 53170-53190, 53650-53651, 53653-53655,53682, 53683	36,000
072	HVE Information technology	53652, 56135	37,000
073	HVE Chemical technology	53670-53676	7,000

No.	Educational type	SOI-code	Number of workers av. 1992, 1994
074	HVE Transport and logistics	54100-54999	28,000
	HVE Paramedical services		
075	HVE Nursing	55120, 55380	44,000
076	HVE (Physio)therapy	55140-55149, 55151 55153, 55180, 55238	36,000
077	HVE Nutrition	55420, 58180	5,000
078	HVE Radiology	55241-55243	7,000
079	HVE Other paramedical services	55110-55119, 55130-55133, 55160, 55190-55199, 55231 55248, 55280, 55480	6,000
	HVE Economics		
080	HVE Accounting and business administration	56110, 56133, 56138, 56210 56240, 56249, 56630	54,000
081	HVE Commerce	56143-56180	40,000
082	HVE Tourism and recreation	56142	8,000
083	HVE Legal and fiscal	56611-56629, 56690	22,000
084	HVE Secretarial	55310, 56131	38,000
085	HVE Business administration technology	52310, 52350, 52380, 56120, 56220	46,000
	HVE Social and cultural		
086	HVE Communication and journalism	56141, 57120	11,000
087	HVE Social and welfare work	51610, 57111, 57115-57119 57188-57199	65,000
088	HVE Personnel work	57113	20,000
089	HVE Library and documentation	57130	15,000
090	HVE Other social and cultural	57141, 57143, 57145-57165	11,000
	HVE Fine arts		
091	HVE Performing and visual arts	58610-58640, 58680-58699, 68610	45,000
	HVE Public order and safety		
092	HVE Public order and safety	59100-59499	10,000

No.	Educational type	SOI-code	Number of workers av. 1992, 1994
	HVE Other		
093	HVE Other	53685-53699, 53868, 53885, 53900, 56190-56199, 58110-58125, 58190, 58340, 58900-58999, 59900-59999	19,000
	UE		
	UE Arts and theology		
094	UE Arts	61100-61199, 71100-71199	42,000
095	UE Theology	61610, 71610	6,000
	UE Agriculture and environmental science		
096	UE Agriculture and environmental science	62110-62130, 62158 62170-62180, 62182-62189, 62200-62999, 72175	11,000
	UE Technology		
097	UE Mathematics and natural sciences	63110-63199, 63631, 63670-63688, 73110-73199, 73674, 73675	63,000
098	UE Construction	63613-63618, 63818, 73618	11,000
099	UE Civil engineering	63623-63625, 73623	6,000
100	UE Mechanical engineering	62181, 63643-63649, 73645-73648	10,000
101	UE Electrical engineering	63650-63651, 63653-63659, 72181, 73650-73651, 73653-73659	11,000
102	UE Information technology	63652, 66135, 73652, 76135	7,000
	UE Medical		
103	UE Veterinary and medical sciences	65111, 65160, 75111, 75160	44,000
104	UE Dentistry	65131, 75131	6,000
105	UE Pharmacy	65200-65299, 75200-75299	5,000
	UE Economics		
106	UE Econom(etr)ics	66110, 66141-66219, 66221-66240, 76110, 76141-76219	32,000

No.	Educational type	SOI-code	Number of workers av. 1992, 1994
107	UE Management	66120, 66220, 76120	11,000
108	UE Accounting and fiscal science	66133, 66630, 76133, 76630	15,000
109	UE Law and public administration	66118, 66610-66629, 66631-66699, 76118, 76610-76629	50,000
	UE Social and cultural		
110	UE Social sciences	65148, . 65180-65199, 67120, 67130, 67141-67188, 68110, 68180, 75148, 75151, 75180, 77130, 77141-77188	75,000
111	UE Other social and cultural	67110-67119, 67190, 77113	11,000
	UE Fine arts		
112	UE Fine arts	68640	--
	UE Other		
113	UE Other	63690, 63900, 65380, 65480, 65900, 68690, 69150-69999	--

PVE = Preparatory Vocational Education
LGSE = Lower General Secondary Education
HGSE = Higher General Secondary Education
IVE = Intermediate Vocational Education
HVE = Higher Vocational Education
UE = University Education