

A SECTORAL TAXONOMY OF EDUCATIONAL INTENSITY

STATISTICAL CLUSTER ANALYSIS AND VALIDATION

MICHAEL PENEDER*

Word count:

9.412 words (excl. abstract, footnotes, appendix, references, tables, figures, formulas);
+ abstract: 91 words;
+ footnotes: 454 words;
+ appendices/supplement: 991 words
+ references: 882 words
= 11830 words in total.

Corresponding author:

Michael Peneder
Austrian Institute of Economic Research
(WIFO)
P.O. Box 91; A-1030 Vienna
Austria
Tel.: + 43-1-798 26 01 – 480
Fax.: + 43-1-798 93 86
E-mail: michael.peneder@wifo.ac.at

* The author gratefully acknowledges support from the European Commission's Fifth framework programme, under the project, Employment Prospects and the Knowledge Economy, HPSE-CT-2001 00055.

Abstract

The paper classifies forty-nine manufacturing and service industries according to their educational workforce composition. Statistical cluster techniques are applied to data for the USA, Germany, France, the UK and Austria. Industries are first classified separately for each country, providing an appropriate tool for the analysis of national micro-data. Later, we propose a common consensus classification, enabling comparative international studies. Validation of the cluster solution reveals considerable robustness to variations over time and between countries. Finally, we test for statistical associations with sector performance, confirming a significant tendency towards ‘education-biased structural change’.

Key words: industry classification, taxonomy, human resources, education, labor skills

JEL Codes: C19, I20, J21, J24, L80

1. Introduction

1.1 Motivation and outline

Apart from the cultural values of education to the individual and society at large, the economic interpretation of education emphasizes its nature as a special input to production. This 'human capital revolution' (Freeman, 1986) was triggered by Theodore W. Schultz (1960, 1961), who proposed that education is an investment in people that generates a distinct class of productive assets – labeled *human capital*, because it distinctly becomes part of the person receiving it. Schultz also pointed at the particular importance of human resources in the process of economic development and presented empirical evidence for the continual rise of this investment. Although such evidence is naturally based on the individual characteristics of the people occupying the jobs, it also reflects the labor skill requirements of firms, which in part depend on the characteristics of markets and industries.

The paper focuses on this aspect of sector-specificity in the demand for human resources, presenting a new sectoral taxonomy of educational intensity. We use international data on educational attainment and apply statistical cluster techniques to identify up to 49 manufacturing and service industries in mutually exclusive classes. The introductory section clarifies the motivation and presents basic conceptual considerations. Section 2 provides a brief methodological commentary on statistical cluster techniques and explains the particular choices made during the course of the analysis. Section 3 explains the data we used. Section 4 begins with a detailed documentation of the clustering process for the USA, which serves as an illustrative example of analogous work for the UK, France, Germany and Austria. We then propose an international classification and test for the general validity of the final cluster identification with respect to variations over time and between countries. Section 5 investigates the new taxonomy's discriminatory power with respect to growth of value added, employment, labor productivity, and other variables. Section 6 summarizes and concludes.

Before moving on to the statistical exercises, we want to address two fundamental queries: (i) 'why education?'; and (ii) 'why a sectoral taxonomy?'. The first question concerns the individual incentives to invest in schooling, and thus the supply of skills, establishing the causal link between education, earnings and productive capabilities. The second query draws our attention towards the demand side, in particular addressing the causes of sector specificity in labor-skill requirements. Finally, we conclude this introductory section with a discussion of the data used in the subsequent cluster analysis.

1.2 The supply of labor-skills

Requiring the devotion of substantial resources, *human capital theory* regards education as an investment, which improves a person's productive capabilities and hence his or her individual prospects for future earnings. Gary S. Becker (1964/75) has been the first to offer a systematic and consistent analysis of the individual choices to invest in training and schooling. Based on the rather strong assumptions of perfectly competitive factor and product markets, he maintained that an informed and rational individual invests in education as long as the expected rate of return exceeds the cost, which is comprised of direct expenses on education plus foregone earning opportunities. Under proper conditions, each individual attains an optimal level of education, which in models of perfectly competitive markets equals the social optimum.

Signaling theory (Spence, 1973, 2002; Arrow, 1973) offers an alternative explanation of the schooling-earnings link. According to this view, education primarily solves the problem of asymmetric information, by distinguishing between low- and high-productivity personnel. It is assumed that at the time of hiring, only the individual is familiar with his or her productive capabilities, whereas the employer is not. Learning about them takes time and therefore incurs costs. In this situation, it is rational for an individual to acquire academic credentials, even if it is for the sole purpose of signaling one's capability to do so. Educational attainment is a credible signal as long as the cost of schooling and productivity are negatively correlated. This is a plausible assumption, when, e.g., people of greater ability tend to experience lower cost of educations (because they learn faster) or expect higher returns (because they learn better), in both cases raising their optimal levels of schooling.

Signaling theory applies the same individual rationality to the choice of schooling as the human capital model does. The crucial difference is that people rationally invest in schooling, even though school might not enhance their productive capabilities at all. The signaling mechanism itself is productive through its matching of people with heterogeneous abilities to appropriate jobs. But, the private and social returns to education differ, because part of the individual cost only affects the distribution but not the generation of income and therefore the model generally predicts overinvestment in education.

An opposite view emerges, if we take the *positive spillovers* of education and training into account. Sianesi and Van Reenen (2002, p. 5f) mention a variety of potential channels for such external effects. For example, an increase in human capital can (i) raise the speed of innovation and the pace at which new technologies, work practices, etc. are adopted; (ii) increase the incidence of learning from others; (iii) stimulate the accumulation of complementary physical capital; and finally (iv) improve other

social conditions (e.g., public health, parenting, lower crime, contribute to a healthier environment, increase political participation and social cohesion) with a positive impact on productivity. All this suggests that the social returns of schooling exceed the private returns and individuals are therefore likely to underinvest in education.¹

The upshot is, that despite large conceptual differences, all relevant economic models treat schooling as an investment in future earnings. They reveal three channels of how education positively affects economic development: first, through the acquisition of cognitive and social skills (human capital theory); second, by sorting high- and low-productivity personnel into appropriate jobs (signaling and screening); and third, by increasing a society's capacity for innovation and the diffusion of new ideas (positive spillovers). All three mechanisms support the conclusion that educational intensity is a valid measure of the productive capabilities available in the human resource base of a firm, sector or country.

Still, we must keep in mind important caveats that arise from our exclusive focus on educational attainment. First, formal education is only one of several elements that comprise the productive capabilities of an employee. A more complete measure of 'skill-intensity' would also contain information about experience, on-the-job training, and unobserved capabilities that are not appropriately reflected in schooling credentials. The apparent reason for not considering them in this paper is the lack of reliable and comparable international data. Secondly, our measures of educational intensity cannot account for variations in the quality of schooling, i.e. when different institutions offer equivalent formal degrees.² Finally, a purely economic interpretation falls short of understanding the psychology and social conditionality of schooling choices.³

¹ See, for example, Dearden, Reed, and Van Reenen (2000), or Sianesi and Van Reenen (2002). De la Fuente and Doménech (2002) report a strong and positive relationship between data quality on the one hand, and the size and significance of human capital coefficients in growth regressions on the other.

² See Card and Krueger (1992), or Brewer, Eide and Ehrenberg (1999).

³ For instance, Akerlof and Kranton (2000, 2002) integrate the psychological concept of 'identity' into economic models of demand for education as a separate argument in utility functions.

1.3 The demand for labor-skills

Assuming that factor and product markets are perfectly competitive, the most straightforward explanation of variations in the demand for educated personnel are intrinsic differences in the technology of production, which determine the marginal product, and together with input prices the factor shares of distinct skill classes.⁴ For a given level of output, the respective ratio of wages to labor productivity is therefore the immediate criterion in selecting skill standards for heterogeneous types of labor. Lazear (1998, p. 23ff) further specifies three circumstances under which this general principle has to be amended, if it is to encompass the true price of labor and its respective impact on output. Two of them relate to technology and one to the labor market. The first case is when production requires a large deal of interaction with other complementary assets, such as expensive machinery, as well as equipment, patents or brands which have intangible sunk costs in R&D and advertising. The reason is that when labor is priced accurately, the price also includes the opportunity costs of other assets which may be tied up, in addition to wages directly accrued. Secondly, the demand for skilled relative to unskilled labor tends to rise, when work practices require a high degree of interaction among the company's personnel, because the contribution of any employee to overall output also includes an effect on his or her co-workers' output. Finally, when the labor market is 'thin' and, for instance, persons with very specialized skills are difficult to find, highly specialized personnel will earn higher wages, because employers also include the additional costs of searching in their calculations.

In short, from the micro-perspective of a human resource manager, the required skill standards largely depend on the characteristics of the technology and labor markets. In principle, both are exogenous to the firm, once it has chosen its geographic location and product portfolio. We should therefore expect that these determinants correlate with sector-specific contexts, as defined, for example, in standard industry classifications. The general hypothesis then states that in addition to any direct intrinsic differences, an industry's educational intensity tends to rise with the degree of interaction within its workforce, the capital intensity of production, and the 'narrowness' of the labor market with respect to specific skills.

⁴ See Hamermesh (1993) for an elaborate discussion of the demand for heterogenous labor under a variety of production and cost functions.

Shifting our focus from firm-level choices to macro trends, four particular aspects of economic development appear to have a lasting influence on the demand for educated labor: (i) the income elasticity of demand for knowledge intensive goods and services; (ii) growing international trade; (iii) technological and organizational change; and (iv) additional feedback from an increasing supply of skilled-labor. First, the income elasticity of demand is generally believed to be high for immaterial sources of well-being. As the private consumption of material goods is more quickly affected by the saturation of markets, especially in the case of physiological limits to further consumption (e.g. foods and tobacco), knowledge-intensive goods and services tend to face better opportunities for raising demand and output in correspondance with increases in disposable income per capita.

Second, in an open economy, *international trade* affects product prices and thus directs the composition of a firm's product portfolio towards industries with (economy-wide) comparative advantages. According to standard Heckscher-Ohlin trade theory, countries, which are relatively abundant in skilled labor, will therefore specialize in skill intensive industries. As a consequence, in developed economies the growing integration of world markets through trade liberalization, as well as falling transport and communication costs, tend to raise the relative demand for skilled personnel and dampen the wages of unskilled personnel. Although the precise strategy and quantitative impacts are disputed, the principle mechanism is supported by empirical data.⁵

Third, in recent decades the tremendous progress in computer technologies has boosted the academic interest in the influence of *technological change* on skill requirements.⁶ Caselli (1999) explains that the adoption of new technologies leads to substitution between different kinds of capital, hence affecting skill requirements. Operating new machines and equipment can either become easier to learn ('de-skilling'), or more difficult, requiring better education and training ('up-grading'). In other words, the effect of new capital on skill requirements depends on the particular technology and can go in both directions. An often cited example of 'de-skilling' is the introduction of assembly lines, e.g., in the production of automobiles, which substituted low-skilled factory workers for high-skilled craftsmen. In contrast, the introduction of new computer technologies is the standard example of a positively 'skill-biased technological change' (SBTC), where the technology-induced growth in demand for better

⁵ See, e.g., Wolff (2003), or Forbes (2001).

⁶ See, for example, Krueger (1993), Autor, Katz and Krueger (1998), Falk and Seim (2001), or Chun (2003).

educated personnel accounts for a simultaneous rise in both the employment and wages of skilled labor.

Quoting more evidence on SBTC in general, Machin and Van Reenen (1998, p. 1232) demonstrate "a significant association between skill upgrading and R&D intensity", whereas Kahn and Lim (1998) find a positive relationship between skill intensity and TFP growth. Most relevant to our work are the findings of Berman, Bound and Machin (1998), which show that among developed countries, the within-sector upgrading of labor skills has been largely concentrated in the same industries, while Haskel and Slaughter (2002) additionally claim that SBTC itself has a sector bias. Using international data for 10 OECD countries over the 1970s and 1980s, they demonstrate that rising skill premiums are often caused by technological change that is concentrated in skill-intensive sectors, and vice versa.

Although closely related to technological progress, the impact of *organizational change* is regularly neglected, due to the lack of adequate data. However, two recent empirical studies managed to break this barrier and found evidence of 'skill-biased organizational change' (SBOC) – defined by the complementarity between increasing labor skills and changes in modern work practices. Investigating British and French micro-data, Caroli and Van Reenen (2001) explain that the recent move towards the decentralization of authority and shorter chains of command depend on the availability of educated personnel, who can handle increased responsibility and are more likely to enjoy this kind of job enrichment. Of related concern, Bresnahan, Brynjolfsson and Hitt (2002) use data of large US companies to demonstrate that the demand for labor skills is embedded in a three-way system of IT-capital, new services, and new organizational practices, in which each complements the other.

Finally, Acemoglu (1998) cites the increasing *supply of labor skills* as a third economy-wide force which systematically shifts skill requirements upward. His model points out that technological change responds to profit opportunities from changing factor prices. Technology and labor skills thus become endogenous and SBTC is explained by the relative abundance of skilled labor in developed countries.⁷ Adding the dimension of international trade strengthens this mechanism to a still greater extent.⁸

⁷ "The twentieth century has been characterized by skill-biased technical change because the rapid increase in the supply of skilled workers has *induced* the development of skill-complementary technologies" (Acemoglu, 2002, p. 9).

⁸ In the model designed by Acemoglu and Ziliboti (2001), the less developed countries suffer from the mismatch of a low-skill workforce on the one hand, with imported technologies developed for high-skill production on the

In summary, during the past decades factor markets, technology, organization, and trade interdependently caused the demand for labor skills to grow. At a given level of output, technology and wages are the immediate determinants of the skill standards that individual firms set when they hire employees. But the relative abundance of educated labor also directs research effort towards skill intensive technologies. International trade strengthens this mechanism through price effects and increasing specialization. At the same time, complementary organizational change also fosters the demand for educated labor. Among these various influences, differences in factor markets primarily invoke country specific effects (e.g., due to variations in the educational systems), whereas the technology and trade based explanations suggest a significant sector dependent component. Together with the empirical observation of strong concentration in skill upgrading across industries, these considerations establish a firm foundation for attempting a sectoral taxonomy.

In actual data, however, one will always find large variations within industries. This ought not discourage our effort, as long as we remain aware of at least three causative factors. First, there is considerable heterogeneity of goods (and hence technology) within the statistically defined sectors. Second, most companies are multi-product firms that differ in the portfolios of goods they produce and sell. Third, the literature on organizational behavior generally points towards substantial firm level differences. Since product and factor markets are almost never perfectly competitive, various combinations of technology, strategy, and organizational culture can be viable alternatives, even if firms supply exactly the same product. Or, as Baron and Kreps (1999, p. 38) put it, “[v]ery distinct arrays of policies can fit a given external environment quite well *if* the arrays are internally consistent”.

Despite this undeniable diversity of human resource practices, competitive firms must respond to the constraints of their environment, among them location dependent aspects of relative factor abundance, as well as the industry specific context of technology and markets (Peneder, 2001, 2003). Focusing on the latter, the new taxonomy is designed to capture a significant part of that variation.

other. By reducing the prices for unskilled production in the developed countries, trade discourages R&D in unskilled technologies even further. As a consequence, international trade tends to increase the skill-bias of technological change.

2. The data

The most sensitive choice in any statistical cluster analysis is the initial decision concerning the appropriate dimensions against which individual cases should be measured and discriminated. In addition to an elaborate economic rationale, statistical cluster analysis benefits from a clear concept of geometric space that allows for a meaningful measurement of distances between observations. We must choose the variables in a way that spans the independent dimensions of the phenomenon under investigation.

Regarding the economic rationale, the empirical observation "that more-highly educated and skilled persons almost always tend to earn more than others" has been cited by Becker (1964/75, p. 10) as "the most impressive piece of evidence" in favor of the human capital approach. Signaling theories assume as well that the prospect of raising future earnings is the basic incentive for investing in additional years of schooling. Since then, numerous econometric studies have confirmed the positive school-to-earnings link (see Card, 1999). With this literature in mind, we can take it for granted that educational attainment is not just a valid ordinal measure of schooling inputs (since higher degrees require more years and therefore more investment in human capital and reputation), but also correlates positively with the productive capabilities in the human resource bases of firms. The ordinal relationship between levels of educational attainment allows us to map the different patterns of schooling credentials into a single dimension of educational intensity, used to classify industries into separate categories ranging from 'very low' to 'very high'.

In the current procedure, an industry's workforce is segregated by the individual's highest level of educational attainment, for which, depending on the availability of data, the shares in total employment, wages or hours worked were calculated. The shares of all but one of these educational categories were then entered in the clustering algorithm. Since each person is a member of only one of the various educational categories, the latter effectively span an orthonormal space for each measure of workforce composition (i.e. employment, wages or hours worked). However, taking benefit from the additional information that we gain by adding the shares in total wages to those of employment, when available, both variables were used. Because of their almost perfect symmetry, their lack of independence does not adversely affect the clustering process. Taking into account the time dimension is of related interest but regards the independence of observations and not of variables. A special aim of the national taxonomies is to map the evolution of industries as they move through different skill classes over time. Therefore each year is added as a set of independent observations so that the total number of cases is given by the number of industries multiplied by the number of years. Although this

introduces a high degree of correlation between observations, it also affects the formation of clusters nearly symmetrically for all industries. Finally, we choose not to standardize the data, first, because all variables are already reported in the same units (e.g. employment shares), and second, because of a desirable implicit weighting of educational categories in proportion to their overall variation between industries.

The data were collected in a collaborative effort, which was funded under the 5th Framework Programme of the European Commission and coordinated by the National Institute of Economic and Social Research (NIESR) in London. Other contributing institutions that helped to aggregate national micro-data into a common sectoral breakdown are The Conference Board in New York, the Groningen Growth and Development Center (GGDR) in the Netherlands, the Centre d'Etudes Prospectives et d'Information Internationaux (CEPII) in Paris, the Zentrum für Europäische Wirtschaftsforschung (ZEW) in Mannheim, and the Austrian Institute of Economic Research (WIFO) in Vienna. The data are for the USA, Germany, France, the UK and Austria and (with considerable variations between countries) cover the period from 1979 to 2000. The annual data were pooled to comprise two consecutive years. Except for Germany, all the data were extracted from national labor force surveys, which are based on large-scale household interviews of individuals, and include information on educational qualifications and occupations (of employees and the self-employed) and also the codes of the industries where the people work. The German data are based on the official employment statistics instead of the labor force surveys. It offers a complete count of employees according to occupations and educational attainment, but does not include the self-employed.

Table 1 summarizes relevant information regarding the data, which is used in the statistical cluster analysis. Despite a remarkable attempt by NIESR to bridge the differences between the national data sets, many inconsistencies have remained with respect to the precise meaning and measurement of the variables. For instance, the composition of the workforce according to educational attainment was measured by three different variables: (i) the shares in total employment (all countries except France), (ii) total wages (USA and UK), and (iii) total hours worked (France). Furthermore, it was not possible to match the different national nomenclatures perfectly into one common and comparable industry list.

Table 1: Summary of the data used in the cluster analysis

Country (data source)	Time, sectors and sample size	Measure of composition	Measure of educational attainment
USA (Current Population Survey)	Annual data from 1979 to 2000; 39 sectors; 50.000 households each month;	Share in total - employment; - wages;	1. University degree 2. Associate degree 3. Some college 4. High school education 5. No formal qualifications
Germany (Beschäftigtenstatistik)	Annual data from 1988 to 1998; 40 sectors; Total population of employees (no self-employed) per quarter;	Share in total - employment	1. University degree 2. Technical university ("Fachhochschule") 3. Vocational + university entrance degree 4. University entrance degree 5. Vocational + basic or intermediate general education 6. No formal qualifications
France (Enquête Emploi)	Annual data from 1989 to 1999; 34 sectors; 75.000 households once a year;	Share in total - hours worked; - wages;	1. Baccalauréat + 4 years (university or specialised school) 2. Baccalauréat + 2 years (university or specialised school) 3. High school finishing degree ("baccalaureat") 4. Vocational high school 5. High school entrance degree (after 4 th year of secondary education) 6. No formal qualifications
United Kingdom (Labour Force Survey)	Annual data from 1979 to 2000; 41 sectors; 60.000 households per quarter;	Share in total - employment; - wages;	1. University degree 2. Higher education below degree level 3. Intermediate vocational (generally technical) qualifications 4. High school education 5. No formal qualifications
Austria (Mikrozensus)	Annual data from 1995 to 2000; 46 sectors; 33.000 households per quarter;	Share in total - employment	1. University degree 2. Technical university ("Fachhochschule") 3. Vocational + university entrance degree 4. University entrance degree 5. Vocational + basic or intermediate general education 6. No formal qualifications

The major problem, however, is that the measures of education lack comparability. The profound differences between countries in the organization of educational systems are for real and cannot be reasonably reduced to nominal problems of nomenclature or different labels and aggregations. Consequently, as Kerckhoff and Dylan (1999) have demonstrated, the harmonization of indigenous credentials into standard categories "tends to distort some of the countries' similarities and differences" (ibid., p. 762). Contrary to the initial plan to pool the sectoral information of all countries, we decided to refrain from fitting the data into standard categories of educational attainment. As a

consequence, the statistical cluster analysis was applied to each country separately. Instead of a single international industry classification, we first produced five country specific taxonomies and then synthesized these into one international ‘consensus’ classification.

3. Statistical cluster analysis

3.1 General aim

Statistical cluster analysis is defined as “the art of finding groups in data“ (Kaufmann and Rousseuw, 1990) such that the degree of “natural association“ (Anderberg, 1973) is (i) high among members within the same class (*internal cohesion*) and (ii) low between members of different categories (*external isolation*). In practice, internal cohesion and external separation are not definite requirements, but rather general objectives. Their fulfillment is a matter of degree and depends on the nature of the data as well as the clustering techniques applied.

Cluster analysis offers a sophisticated statistical tool for the exploration and classification of multivariate data, but it is important to acknowledge that it remains a heuristic method, which requires the researcher to make a number of choices that critically affect the final outcomes. In the following, we present a brief explanation of relevant techniques in chronological order of the deliberate choices made in the current analysis.

3.2 Measures of (dis)similarity

Once the variables are chosen, the clustering procedure starts with a given data matrix of $i = 1, \dots, n$ observations for which characteristic attributes x are reported for $j = 1, \dots, p$ variables. The initial data set of the dimension $n \times p$ is then transformed into a symmetric (dis)similarity matrix of dimensions $n \times n$ observations with d_{ih} being the coefficients of (dis)similarity for observations x_i and x_h .

$$(1) \quad D_{n,n} = \begin{bmatrix} 0 & \dots & & & 0 \\ d_{21} & 0 & \dots & & \\ d_{31} & d_{32} & 0 & \dots & \\ \vdots & & \vdots & & \\ & \dots & d_{ih} & \dots & \\ & & \vdots & & \\ d_{n1} & d_{n2} & \dots & d_{n(n-1)} & 0 \end{bmatrix}$$

For any observations x_i , x_h and x_g with i, h , and $g = 1, \dots, n$, located within measurement space \mathbf{E} , the desired formal properties of the (dis)similarity matrix \mathbf{D}_{nn} are defined as follows (Anderberg, 1973, p. 99):

1. $d_{ih} = 0$ if and only if $x_i = x_h$, i.e. for all observations the distance from itself is zero and any two observations with zero distance are identical;
2. $d_{ih} \geq 0$, i.e. all distances are non-negative;
3. $d_{ih} = d_{hi}$, i.e. all distances are symmetric; and finally
4. $d_{ih} \leq d_{ig} + d_{hg}$, known as the triangle inequality, which states that going directly from x_i to x_h is shorter than making a detour over object x_g .

The combination of the first and second properties assures that \mathbf{D}_{nn} is fully specified by its values in the lower triangle. The fourth property establishes that \mathbf{E} is an Euclidean space and that we can correctly interpret distances by applying elementary geometry. Any dissimilarity function that fulfills the above four conditions is said to be a *metric*.

In this spirit, the *Euclidean distance* e_{ih} appears to be the most natural measure of (dis)similarity, due to its direct application of the Pythagorean theorem, which states that the hypotenuse of a right triangle is equal to the square root of the sum of the squares of the other two sides:

$$(2) \quad euc_{ih} = \sqrt{\sum_{j=1}^p (x_{ij} - x_{hj})^2} \quad 0 \leq euc_{ih} < \infty$$

Operating with the squared differences, the Euclidean measure will, for example, rank two observations with a difference of 1 unit in the first variable and 3 units in the second variable as farther apart than two observations with a difference of 2 units in both variables. In other words, it is sensitive to outliers. Alternatively, the closely related Manhattan or *city block distance* prescribes equal importance to any unit of dissimilarity, because it simply calculates the sum of the absolute lengths of the other two sides of the triangle:

$$(3) \quad cityb_{ih} = \sum_{j=1}^p |x_{ij} - x_{hj}| \quad 0 \leq cityb_{ih} < \infty$$

Kaufmann and Rousseeuw (1990, p. 12) use the image of a city in which the streets run vertically and horizontally to explain the peculiar name. The Euclidean measure corresponds to the shortest geometric distance a bird could fly straight from point x_i to point x_h , whereas the use of the Manhattan

measure is consistent with the distance that people would have to walk around the city blocks. Both measures in (2) and (3) fulfill the requirements of a metric.

When we are interested in the ‘shape’ of objects rather than in the absolute size of differences, alternative measures can be more helpful. The following two measures of similarity, called *angular separation* in (4) and the *correlation coefficient* in (5), are most frequently used:

$$(4) \quad ang_{ih} = \frac{\sum_{j=1}^p x_{ij}x_{hj}}{\sqrt{\sum_{j=1}^p x_{ij}^2 \sum_{j=1}^p x_{hj}^2}} \quad -1,0 \leq ang_{ih} \leq 1,0$$

$$(5) \quad corr_{ih} = \frac{\sum_{j=1}^p x_{ij}x_{hj} - (1/p)(\sum_{j=1}^p x_{ij} \sum_{j=1}^p x_{hj})}{\sqrt{\left[\sum_{j=1}^p x_{ij}^2 - (1/p)(\sum_{j=1}^p x_{ij})^2 \right] \left[\sum_{j=1}^p x_{hj}^2 - (1/p)(\sum_{j=1}^p x_{hj})^2 \right]}} \quad -1,0 \leq corr_{ih} \leq 1,0$$

Both angular separation and the correlation coefficient measure the cosine of the angle between two vectors. The essential difference between the two is that the former is based on deviations from the origin, whereas the latter operates with deviations from the mean of the variables of an observation. As a consequence, the correlation coefficient is unaffected by mere size displacements (i.e. the uniform addition of a constant to each element). The correlation coefficient is therefore less discriminating than the angular separation measure.⁹

In Appendix A we provide a simple numerical example plus geometric visualisation that demonstrates, why the choice of various measures has an apparent impact on the values of the final (dis)similarity matrix \mathbf{D}_{nn} .¹⁰ In some instances, *a priori* conceptual considerations about the nature of the variables and the desired properties of the classification might be a sufficient guide in making that decision. In

⁹ Since correlation-type measures can take negative values, they do not strictly fulfill the above requirements of a metric. Anderberg (1973, p 113f) discusses the “limited metric character“ of the correlation coefficient. However, these measures can be transformed to take values between 0 and 1 by defining $ang_{ih}^* = (1+ang_{ih})/2$ and $corr_{ih}^* = (1+corr_{ih})/2$ (see Gordon, 1999, p. 21).

¹⁰ The literature provides a variety of other (dis)similarity functions that are applied in statistical cluster analysis. For extensive surveys see, for example, Romesberg (1984) and Gordon (1999).

general, however, it is desirable to try out more than one function and to learn how robust the results are with respect to the variations in the concepts of measurement. However, there are also trade-offs to consider and repeatedly increasing the number of (dis)similarity functions inevitably leads to diminishing returns. For the purpose of the current classification, the four measures presented in equations (2) to (5) will provide a reasonable and sufficient range of functions, which allows us to take into account the robustness of the results with regard to different (dis)similarity matrices.

3.3 Clustering algorithms

The next crucial step concerns the choice of how to group objects into separate categories, i.e. we must choose what clustering algorithm to use. Among the clustering algorithms that are most widely used, we must distinguish between two general approaches. The first is the *partitioning* method, which breaks objects into a distinct number of non-overlapping groups. The most common of them, which is also applied here, is the so called *k-means* technique. The second approach is the *hierarchical cluster analysis*, which is either divisive or agglomerative, i.e. dividing or combining hierarchically related objects into clusters. Three variations of the *agglomerative* hierarchical clustering method are used in the current analysis.

For the *k-means* method, the set of observations is divided by a pre-defined number of clusters k . For example, k nearly equal-sized segments can be formed as an initial partition. Cluster centers are computed for each group, which are the vectors of the means of the corresponding values for each variable. The objects are then assigned to the group with the nearest cluster center. After this, the mean of the observations are recomputed and the process is repeated until convergence is reached. This is the case, when no observation moves between groups and all have remained in the same cluster of the previous iteration. With this method, a critical and potentially very manipulative choice is the initial number of clusters k . Outliers in the data can seriously distort the cluster means. By increasing the number k , more and more outliers will be segregated as separate clusters, so that the remaining objects will be classified as though the outlier were not there. But again, there is a trade-off. If the number of clusters k is too large, the problem of missing information about the relative (dis)similarity between clusters makes it difficult to find a meaningful final structure for the total set of observations. We therefore apply the following self-binding rule-of-thumb: “Choose the lowest number k that maximizes the quantity of individual clusters l which include more than 5% of the observed cases“.

In contrast to the *k-means* method, *hierarchical* cluster analysis enables us to determine the boundaries between clusters at different levels of (dis)similarity. Preserving a higher degree of

complexity in the output produced, hierarchical techniques require a heuristic interpretation of the surfacing patterns. Dendrograms (or ‘cluster trees’) support this by means of graphical representation. The branches on the bottom of the chart represent one entity each, while the root on top represents the entire set of objects. As we move upwards on the chart, the degree of association between objects is higher, the sooner they are connected by a common root. Conversely, objects or groups are the more dissimilar, the longer they remain disconnected. As with k-means cluster analysis, any of the above measures of distance can be applied. When groups with more than one object merge, various methods differ in the way they determine what the (dis)similarity between groups precisely is. The most popular and intuitively appealing choice is the *average linkage* method, whereby the average (dis)similarity between all the observations is compared for any pair of groups. Alternatively, the *complete linkage* method compares the (dis)similarity between the observations which are farthest apart, whereas the *single linkage* method takes the (dis)similarity of the nearest neighbors in any pair of groups into account.

The choice between the different linkage methods directly relates to the objectives of internal cohesion and the external isolation of clusters (Gordon, 1999, p. 84ff). Single linkage aim only for *external isolation*, implying that any observation is more similar to some other object within the same cluster than to any other objects outside. Due to this property, single linkage methods frequently fail to reveal much structure within the data. The reason is that observations tend to join one common and expanding cluster, which leads to undesirable ‘chaining’ effects. Conversely, the complete linkage method aims at *internal cohesion*. This leads to compact classes, which, however, need not be externally isolated. The average linkage method avoids both extremes and seeks a compromise between the aims of internal cohesion and external isolation. In the current analysis, we follow Gordon (1999), who recommends that “if it is not possible to determine a single preferred clustering procedure, it is useful to analyze data using two or more ‘sensible’ methods of analysis and synthesize the results. This can involve superimposing classifications onto graphical representations of the data and/or obtaining consensus classifications“ (ibid., p. 100). As he further explains, “the hope is that the results are less likely to be an artifact of a single method of analysis and more likely to provide a reliable summary of any class structure that is present in the data“ (ibid., p. 184).

In the current analysis, we apply a two-step approach that combines k-means and agglomerative hierarchical methods. The k-means method produces a first partition, which reduces the large initial data sets, so they can be used more effectively in the second step of hierarchical clustering. The k-means method also has the advantage that the initial case assignments remain reversible during the

course of iterations. In this first step, we only use the Euclidean measure of distance. For the purpose of further refinement, the resulting cluster centers are redefined as objects for the following hierarchical agglomeration clustering method. This provides us with the advantage of a more detailed hierarchical representation. Longing for a comprehensive evaluation of the association between industries, we apply all of the above four (dis)similarity functions to each of the three agglomeration methods. An inspection of the major regularities in the respective dendrograms then leads to the synthesis of the final classification.

In short, this methodological section has demonstrated the multitude of potentially very influential choices researchers have to make during the clustering process. In order to be trustworthy, any cluster analysis should therefore include a full documentation of the decisions made, together with a detailed explanation of how the graphical representations are interpreted. This is the task of the next section.

4. The taxonomies

4.1 A national example: the US taxonomy

Considering the enormous differences in both the educational and data systems, our initial plan to harmonize the categories of educational attainment in the various countries had to be dropped. Instead we chose a tedious, but more rewarding approach, first creating independent national taxonomies for each country, and then using these to synthesize a common consensus classification. While, for instance, the national classifications are applicable, and indeed preferable, for the analysis of micro-data in the particular countries, the consensus classification establishes an analytic tool for international comparative studies. In the following section we take the USA as an illustrative example, first because it is the largest of the five economies, and secondly because it also appears to be the most reliable in terms of data quality. Analogous analyses have been applied to the four other countries, for which similarly detailed documentations are available upon request.

The data set for the USA is comprised of 39 sectors covering two-year averages from 1979 until 2000, amounting to a total of 429 observations. Following the two-step approach already outlined, the 5% benchmark implies that we first need to determine the lowest number k that maximizes the quantity of individual clusters which include more than 21 cases. Running the k -means algorithm on a dissimilarity matrix made up of Euclidean distances between any pair of observations for all values of k ranging from 2 to 35, the lowest number which fulfils the above rule turns out to be $k = 11$ with $l =$

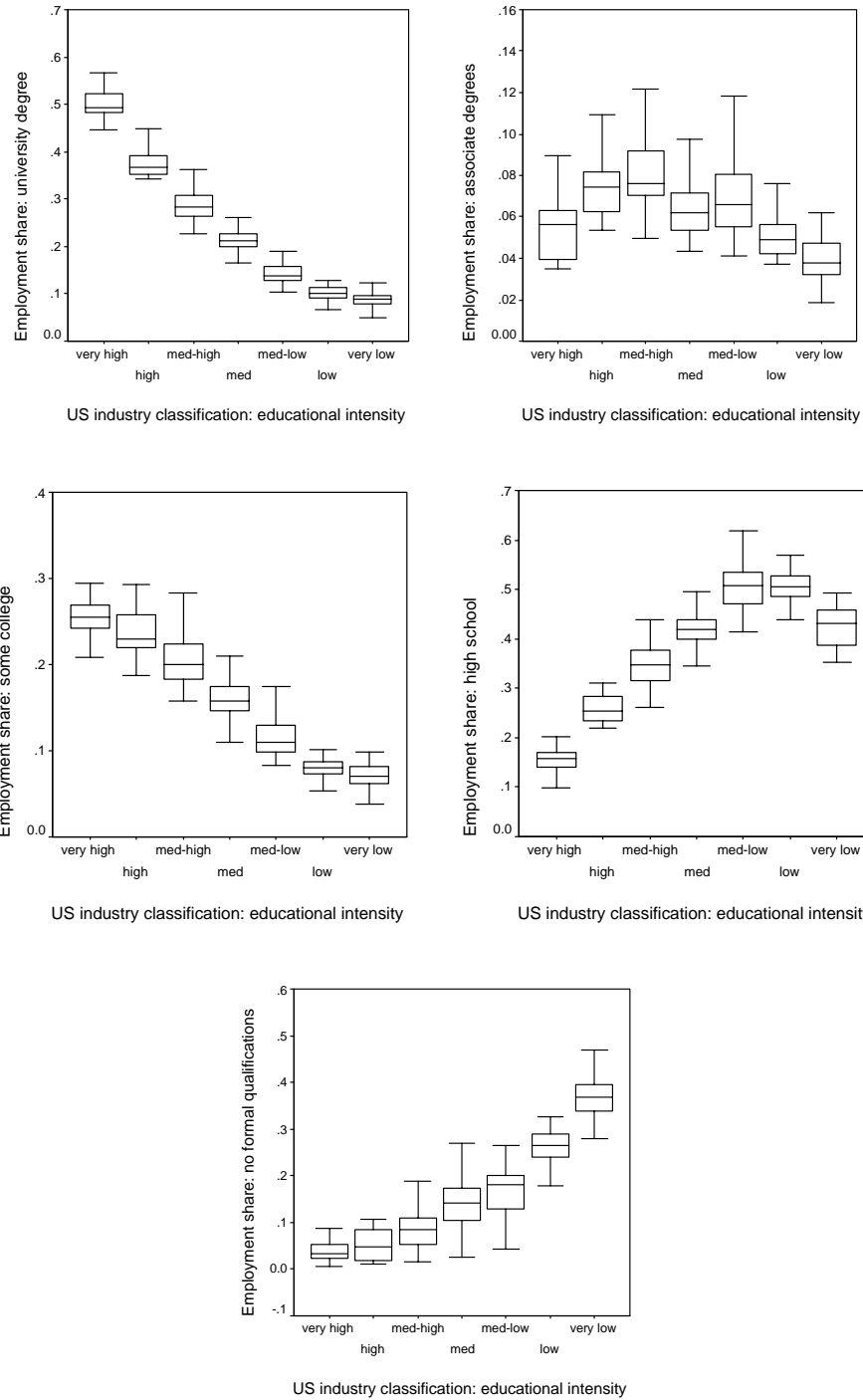
10. This partition has thus produced 11 surprisingly compact clusters, of which only one comprises less than 5% of total observations.

In the second step, the cluster centers from the first partition described above are entered as individual observations in the hierarchical analysis. We use all the four (dis)similarity measures, namely the Euclidean distance (“euc“), the city block distance (“cityb“), angular separation (“ang“), and the correlation coefficient (“corr“), each of which is applied to the three agglomeration algorithms. We present detailed results for the average linkage method (“avl“) in Appendix B, but also produced analogous charts for the complete linkage method (“cpl“) and the single linkage method (“sgl“). All in all, this process established twelve different cluster trees, which we then compared.

Taxonomies should not be seen as an end in themselves; they are created to serve analytic purposes. Their usefulness depends on how well they discriminate between observations with regard to a chosen set of attributes. The boxplot charts in Figure 1 are particularly useful for their *validation*, since they simultaneously display information about the shape and dispersion of educational intensity between various types of industry. The charts are easy to understand. The box itself comprises the middle 50 percent of observations. The line within the box is the median. The lower end of the box signifies the first quartile, while the upper end of the box corresponds to the third quartile. In addition, the lowest and the highest lines outside the box indicate the minimum and maximum values, respectively. The patterns are surprisingly clear, suggesting a pronounced structural (in the sense of industry dependent) dimension of educational requirements. For example, we can easily verify that the share of people with university degrees peaks in the class of industries with very high educational intensity, and then decreases continuously down to the group with very low educational intensity, but is more discriminating in the former than in the latter case.

While the share of people with some college education also decreases continuously from industries with very high educational intensities down to those with very low levels, this variable discriminates best among the intermediate classes of the taxonomy. The share of people with associate degrees is highest in the group of industries with intermediate-to-high educational intensities, whereas persons with some college education or who have graduated from high school are most relevant to the intermediate-to-low category. Finally, the share of people with no formal education discriminates best between the categories ranging from ‘intermediate‘ to ‘very low‘.

Figure 1: Boxplot of employment shares by education and industry type, USA 1989 to 1999



The boxplots highlight two major advantages of multivariate cluster analysis as opposed to conventional "cut-off" methods. The latter usually rely on univariate or bivariate measures, that typically specify a predefined share of sectors (e.g. the 'top ten', or 'top third') as skill intensive industries. In addition to the arbitrary nature of the exogenously defined 'cut-off' line, the exclusive reliance on one variable (e.g., the share of labor with higher education) only enables a satisfactory discrimination at one end of the distribution, while we lose track of the pronounced structural differences at the other end of the spectrum or in the ranges in between. In contrast, statistical cluster analysis endogenously determines the boundaries of the classification from the data and simultaneously processes multiple variables.

Going beyond mere visual validation, the analysis of variance (ANOVA) and simple OLS regressions of educational intensity on the various categories of the industry classification more strictly test the discriminatory power of the new taxonomy. Clearly, the F-statistics, which confirm that for all variables the taxonomy discriminates significantly between observations, are not the issue in Table 2. This result is almost trivial, since the taxonomy was created explicitly for that purpose. More importantly, R-squared and the F-value show us which variables the taxonomy explains more or less successfully. In particular, we can see that the share of between group variation which can be explained by the taxonomy is particularly large for the highest ('university degree') and lowest ('high-school', and 'no formal education') educational categories. In the middle categories, we still find much within group variation in the workforce composition. This directly reflects our experience with the clustering process, where observations located in the middle ranges of the attribute space generally were more difficult to group than those located at either end of the distribution. Furthermore, intermediate qualifications are more dispersed across sectors, because they can substitute personnel with either high or low levels of educational attainment.

Table 2 also presents a convenient way of summarizing average workforce composition, precisely measuring which categories differ significantly and to what extent. The coefficient for the constant term corresponds to the employment share of that educational category in the comparison group (for which we always take the group of industries with a very low educational intensity). In this simple OLS regression, adding or subtracting the coefficients reveals the respective shares in total employment for all other industry types as well.

Finally, we turn to the dynamic properties of the industry-education profile. Although our data span the period 1979 to 2000, the classification itself is invariant with respect to time. As the educational

composition of the workforce changes, industries can traverse the boundaries between types – just as if they were moving upstairs or downstairs on the boxplots.

Apart from minor exceptions, in Table C.1 (in the Appendix C) all industries moved upstairs from lower to higher levels of educational intensity, which is consistent with the aggregate trends. Overall, the temporal patterns in the evolution of educational profiles according to industry are remarkably stable and consistent. This demonstrates another useful application of the taxonomy. Itself being invariant in time, it condenses information from a multivariate attribute space into one single ordinal scale, against which the sectoral evolution of educational intensity becomes transparent.

4.2 The international classification

Data from the four other countries were processed in precisely the same manner as for the US. Overall, the individual cluster analyses worked out very well, producing highly significant and comprehensible separations among the aforementioned categories. But because of substantial differences in the country data concerning attainment levels, measures of composition, as well as the sectors and years covered, each taxonomy had to be created separately. As a consequence, the national taxonomies cannot be compared directly. Although they exhibit a largely corresponding set of categories, ranging from 'very low' to 'very high' educational intensity, each of them is a distinct classification that ought apply only to data of that same country. For the purpose of international comparative analysis, we desire a more representative taxonomy, that is less affected by idiosyncratic country characteristics. Therefore, the following international classification integrates information from the five countries – all of which vary in size as well as industrial and educational systems.

Table 3 summarizes the representation of industries in the five national taxonomies. Ordinal numbers from 1 (very high) to 7 (very low) represent the categories of educational intensity. Since the industry breakdown is not identical across countries, *italic* numbers within parentheses indicate that in the particular country the value is derived from the more highly aggregated sector. For example, the communications sector is comprised of two industries that are likely to differ substantially in their educational requirements. We therefore attempted a more detailed breakdown into 3-digit industry codes. But only the US and the UK provided separate data for 'post and courier activities' (ISIC 641) and 'telecommunications' (ISIC 642). Since they rely on original data, both are in ordinary typeface. In

contrast, the *italic* numbers in parentheses for France, Germany and Austria signal that we inserted a value derived from a corresponding, more highly aggregated sector (ISIC 64).

Table 2: OLS regression on educational intensity, USA 1979 to 2000

Type of industry	University degree β (t)	Associate degree β (t)	Intermediate - vocational β (t)	High - school β (t)	No formal education β (t)
Constant	0.089 (23.47)**	0.039 (15.26)**	0.073 (19.63)**	0.424 (74.25)**	0.375 (55.63)**
Very high	0.410 (67.87)**	0.016 (3.97)**	0.181 (30.57)**	-0.269 (29.64)**	-0.338 (31.65)**
High	0.286 (49.67)**	0.037 (9.59)**	0.165 (29.19)**	-0.166 (19.18)**	-0.322 (31.60)**
Interm./ high	0.196 (43.94)**	0.044 (14.66)**	0.132 (30.17)**	-0.080 (12.01)**	-0.292 (36.90)**
Intermediate	0.122 (25.09)**	0.027 (8.42)**	0.089 (18.58)**	-0.004 (0.57)	-0.234 (27.16)**
Interm./low	0.054 (10.97)**	0.032 (9.61)**	0.042 (8.65)**	0.086 (11.66)**	-0.213 (24.50)**
Low	0.010 (2.04)*	0.012 (3.66)**	0.007 (1.45)	0.084 (11.41)**	-0.112 (13.01)**
Very low	(.)	(.)	(.)	(.)	(.)
Observations	427	427	427	427	427
R-squared	0.96	0.43	0.86	0.88	0.84

Absolute value of t statistics in parentheses; * significant at 5%; ** significant at 1%

The class "very low" educational intensity is the comparison group.

In principle, one could again apply the statistical cluster analysis to create the common consensus classification. Two drawbacks imply that it would be better to do otherwise: First, it is difficult to define appropriate measures of distance for ordinal data without invoking an unwarranted assumption of equidistance between each pair of contiguous categories. Secondly, we feel that computing yet another stage of statistical clustering would ultimately rob the entire process of its transparency. Instead, we identify the common consensus classification simply by taking the median of the five national taxonomies. We first calculate the median without considering the values that were derived from the corresponding, more highly aggregated sector. Only in cases where this median is ambiguous, do we take the next integer literal towards the median for all five countries, i.e. including the *italic* numbers.¹¹

The final taxonomy, which is also listed in Table 3, classifies only the four service sectors – computer related activities (ISIC 71), research and development (73), education (80), and extra-territorial organizations and bodies (99) – as being of *very high* educational intensity. The category *high* educational intensity is also comprised of four sectors: computers and office machinery (30), financial mediation (65), total business services (71-74), and its subgroup of other business activities (74). Among the twelve industries with an *intermediate-to-high* educational profile, we find chemicals (24), instrument engineering (33), and insurance (66). The *intermediate* category is comprised of seventeen industries, among them pulp and paper (21), motor vehicles (34), utilities (40-41), and wholesale trade (51). The *intermediate-to-low* category contains only 5 industries, examples of which are rubber and plastics (25), retail trade (52), or railways (60). Seven industries were classified as *low*, among them food, drink and tobacco (15-16), metal processing (27-28), and construction (45). Finally, seven industries belong to the group of *very low* educational intensity. Examples are agriculture (01-05), textiles and clothing (17-19), and hotels and catering (55).

The boxplots in Figure 2 show that the employment share of people with university degrees peaks in the category of industries with a very high educational intensity, and declines continuously until it

¹¹ The median is preferred to the mean, because the latter would require interval-scaled data. In all but four cases the unambiguous median would be equivalent to the integer of the mean anyway. Furthermore, when individual industries exhibit identical cluster identification, this number is inserted for the higher aggregated sector (e.g., ISIC 17-19 in Germany and Austria). But when the sub-sectors exhibit different cluster identifications, the record of the higher aggregate abides as 'missing' (for instance, ISIC 71-74). For the US, UK and France ISIC 20 and 26 were aggregated together with ISIC 36-37 (miscellaneous manufacturing).

reaches the group of industries with a very low educational profile. We observe the sharpest distinctions on the way down from the highest to the intermediate category, whereas the share of higher education differs little among the remaining lower skill groups. The opposite picture emerges when we look at the employment share of people without any formal degrees. Finally, if we aggregate all other levels of educational attainment as intermediate degrees, we find the highest shares in the intermediate categories of the taxonomy, and they fall continuously towards both ends of the distribution. In other words, the intermediate attainment levels did not discriminate between the highest and the lowest industry types, but they did help to differentiate industries from both ends towards the intermediate categories. Overall, the consensus classification appears to make sense and does not offend any *a priori* expectations about the sectoral distribution of skill-requirements.

In addition to providing a meaningful economic interpretation, industry classifications should be reasonably robust with respect to time and spatial boundaries, especially if they are meant for use in international comparative analysis. Plotting different countries and time periods naturally increases the total variation within industry types. Some countries such as the US generally exhibit higher shares of people with university degrees in each of the categories. Similarly, in any of the industry types these tend to grow over time. But the crucial point is, that notwithstanding the substantial variations between countries and over time, these hardly affect the general patterns of distribution and ranking orders.

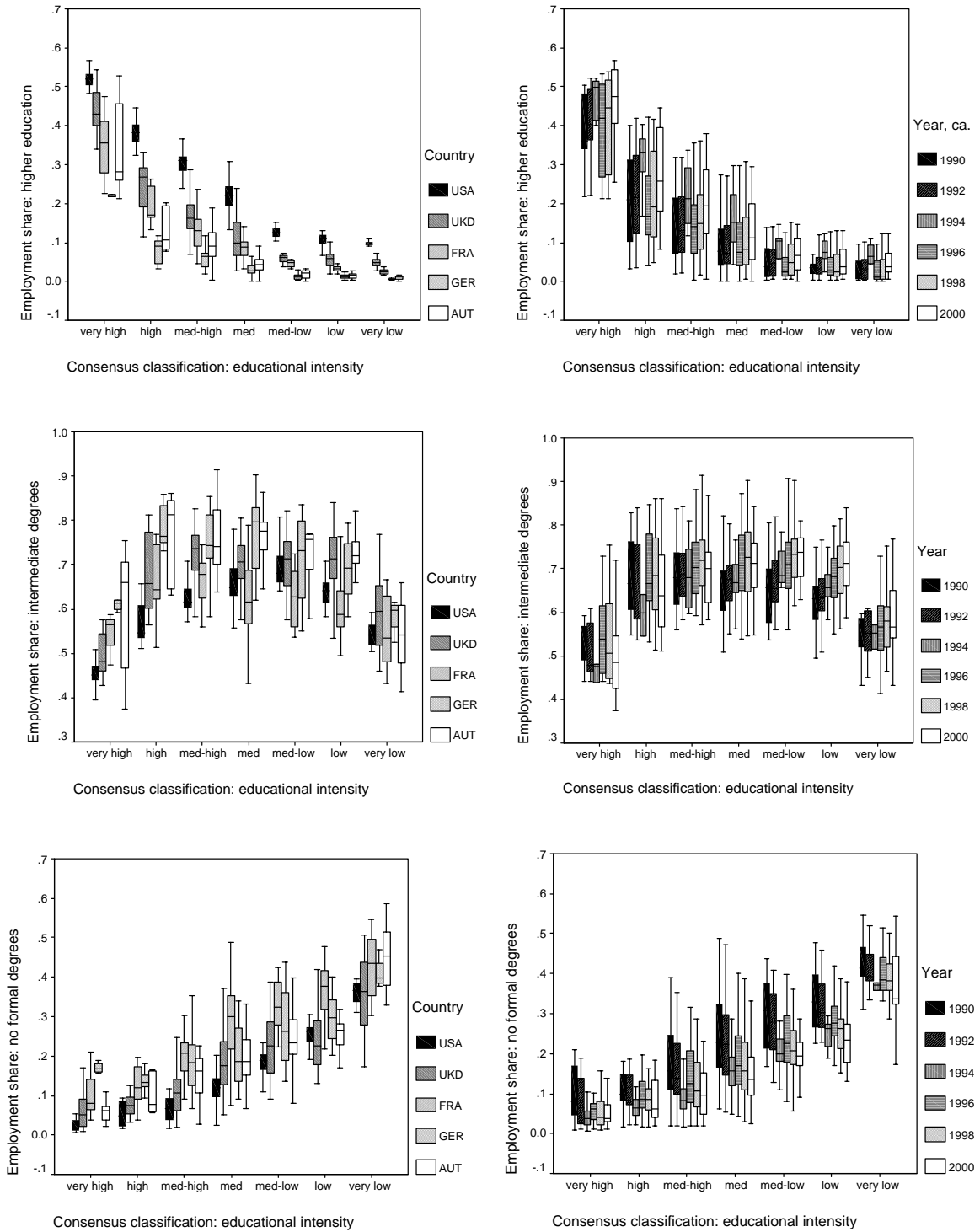
The first piece of quantitative cluster validation is presented in Table C.2 of the Annex. It shows the detailed results of an OLS regressions, where the three variables of educational attainment are explained only by the categorical variables of country, time, and industry type plus interaction effects. Compared to the USA, all of the four other countries exhibit a lower employment share of persons with university degrees, with Germany and Austria exhibiting the largest gap. The UK and Germany have significantly higher shares of intermediate educational attainment levels, whereas France and Austria employ significantly more persons without formal educations. The regression also demonstrates significant time trends, with the share of higher education as well as intermediate degrees continuously rising (i.e. compared to 2000, coefficients are increasingly negative as one goes back in time), while the share of persons without formal educations decreases. Moreover, for each attainment variable, the industry types differ significantly from the comparison group. Although we have now added many additional explanatory variables (including a bulk of interaction effects), the coefficients basically recast the visual patterns of Figure 2 into concrete numbers.

Table 3: The new sectoral taxonomies of educational intensity

ISIC rev. 3	Industry Name	National Taxonomies					International Classification	Notes	
		USA	UKD	FRA	GER	AUT			
01-05	Agriculture, forestry, fishing	7	6	-	7	7	7	very low	
10-14	Mining and quarrying	4	3	6	4	4	4	interm.	
15-16	Food, drink & tobacco	5	5	6	6	6	6	low	
17-19	<i>Textiles, leather, footwear & clothing</i>	7	7	7	7	7	7	very low	
17	Textiles	(7)	(7)	(7)	7	7	7	very low	
18-19	Leather, footwear & clothing	(7)	(7)	(7)	7	7	7	very low	
20	Wood & products of wood and cork	(5)	(5)	(6)	7	(6)	7	very low	*R2
21-22	<i>Pulp, paper products; printing, publishing</i>	4	4	5	3	(6)	4	interm.	
21	Pulp, paper & paper products	(4)	(4)	(5)	(3)	(6)	(4)	interm.	
22	Printing & publishing	(4)	(4)	(5)	(3)	4	4	interm.	
23	Mineral oil refining, coke & nuclear fuel	3	3	3	4	(4)	3	med-high	
24	Chemicals	3	2	3	2	(4)	3	med-high	
25	Rubber & plastics	5	5	5	7	6	5	med-low	*R1
26	Non-metallic mineral products	(5)	(5)	(6)	6	6	6	low	
27-28	<i>Basic metals & fabricated metal products</i>	5	6	6	6	5	6	low	
27	Basic metals	(5)	(6)	(6)	6	5	6	low	*R3
28	Fabricated metal products	(5)	(6)	(6)	6	5	6	low	*R3
29	Mechanical engineering	4	4	6	4	4	4	interm.	
30	Computers, office machinery	1	2	1	2	3	2	high	
31	Electrical machinery & apparatus, nec	3	4	5	4	4	4	interm.	
32	Audiovisual apparatus	2	3	5	4	3	3	med-high	
33	Instrument engineering	3	3	3	3	4	3	med-high	
34	Motor vehicles	4	4	5	3	4	4	interm.	
35	Other transport equipment	3	3	3	4	4	3	med-high	
36-37	Furniture, miscellaneous manuf.; recycling	5	5	6	6	5	5	med-low	
40-41	<i>Electricity, gas & water supply</i>	4	3	3	4	4	4	interm.	
40	Electricity & gas	(4)	(3)	(3)	(4)	4	4	interm.	
41	Water supply	(4)	(3)	(3)	(4)	4	4	interm.	
45	Construction	6	6	6	5	6	6	low	
50	Sale & repair of motor vehicles; retail of fuel	6	6	6	5	5	6	low	
51	Wholesale trade and commission trade	4	5	3	4	4	4	interm.	
52	Retail trade; repair (exc. 50)	4	5	5	5	5	5	med-low	
55	Hotels & catering	7	5	6	7	7	7	very low	*R1
60	Railways & other inland transport	5	5	(6)	5	5	5	med-low	
61	Water transport	5	5	(6)	4	6	5	med-low	
62	Air transport	3	4	(6)	3	2	3	med-high	*R1
63	Auxiliary transport activities; travel agencies	3	5	(6)	3	4	4	interm.	*R3
64	<i>Communications</i>	-	-	2	4	4	4	interm.	*R1
641	Post and courier activities	4	5	(2)	(4)	(4)	4	interm.	*R3
642	Telecommunications	3	3	(2)	(4)	(4)	3	med-high	
65	Financial intermediation (except 66)	2	3	2	3	2	2	high	
66	Insurance and pension funding	2	3	2	3	4	3	med-high	
67	Activities auxiliary to financial intermediation	(2)	3	(2)	(3)	3	3	med-high	
70	Real estate activities	3	2	5	4	7	4	interm.	
71-74	<i>Business services</i>	-	-	-	2	-	2	high	
71	Renting of machinery & equipment	-	5	-	(2)	3	4	interm.	*R2
72	Computer and related activities	1	1	1	(2)	2	1	very high	
73	Research & development	1	1	1	(2)	1	1	very high	
74	Other business activities	2	1	2	(2)	2	2	high	
75	Public admin., defence; social security	2	3	3	3	4	3	med-high	
80	Education	1	1	1	1	1	1	very high	
85	Health, social work	3	2	2	3	3	3	med-high	
90-99	<i>Other Services</i>	4	3	5	-	-	4	interm.	
90-93	Other community, social or personal services	(4)	(3)	(5)	2	3	3	med-high	*R3
95	Private households with employed persons	(4)	(3)	(5)	-	7	7	very low	*R2
99	Extra-territorial organizations and bodies	(4)	(3)	(5)	-	1	1	very high	*R2

Note: 1 = very high; 2 = high; 3 = med-high; 4 = intermediate; 5 = med-low; 6 = low; 7 = very low; numbers in *italics* are derived from the more aggregate data; * decision Rule R1: median overrules mean; R2: unambiguous identification without overrules those including derived values; R3: if identification without derived values is ambiguous, take next integer towards outcome with derived values.

Figure 2: Boxplots by education and industry type with (a) country and (b) time variation



The ANOVA decomposition of variance in Table 4 is another way of assessing the discriminatory power of the new taxonomy. Altogether, country, time, and industry type effects, plus a term for interaction between industry type with countries, explain about 91% of the total variation in the employment share of higher education, 60% of the variation in the employment share of intermediate degrees and 75% of the variation in the employment share of people without formal degrees. Among these variables, the industry types of the new taxonomy are by far the most powerful. Introduced after the country and time effects, they explain an *additional* 53% percent of the total variation in the share of university degrees, as well as of people with no formal education. Only for the share of intermediate degrees, is the unexplained residual (40%) larger than the industry type effect (31%).

Table 4: ANOVA decomposition of total variation

Source of variation	Higher education	Intermediate degrees	No formal education
Sequential Sum of Squares			
Model	13.363	5.348	9.821
Country	4.848	1.738	1.961
Time	0.154	0.172	0.634
Industry type	7.756	2.767	6.889
Ind.*country	0.605	0.671	0.337
Residual	1.252	3.618	3.202
Total	14.615	8.966	13.023
in % of total variation			
Model	91.43	59.65	75.41
Country	33.17	19.38	15.06
Time	1.05	1.92	4.87
Industry type	53.07	30.86	52.90
Ind.*country	4.14	7.48	2.59
Residual	8.57	40.35	24.59
Total	100.00	100.00	100.00

The common consensus classification is necessarily less accurate than the national taxonomies. Similarly, we expect its discriminatory power to decrease with the passage of time. In the following we assess these two kinds of loss of information by means of simple OLS regressions with educational workforce composition as dependent and only the industry classifications as independent variables.

We first use the specific identification of industries for each country in the national taxonomies and then run a second regression with each sector categorized uniformly across countries according to the consensus classification. Table 5 compares the share of explained between group variation to the total variation. Substituting the consensus classification for the national taxonomies, the explained between group variation drops from about 69% to 57% for people with higher education as well as no formal degrees, and from 52% to 37% for intermediate degrees. If we standardize the dependent variable, the explained variation naturally becomes higher. Applying its identification to data for the earliest year available for each country, we can also assess what impact the passage of time has on the accurateness of the classification. Table 5 shows a further drop in explanatory power of only four to eight percentage points. While this finding underscores the need for regular reviews, it also suggests that much change over time remains within the broad boundaries of the given industry types.

In conclusion, the loss of information is substantial but not discouraging. Even after synthesizing the individual country results into one common classification and going back in time up to twenty years, the new taxonomy is able to capture a considerable part of the total variation.

Table 5: Share of variation explained by industry types in total variation (in %)

Classification	Employment share of people with					
	Higher education		Interm. degrees		No formal degrees	
	x	z(x)	x	z(x)	x	z(x)
National taxonomy, latest year	68.72	83.13	52.47	53.53	68.59	69.38
Consensus taxonomy, latest year	57.42	80.29	37.12	41.92	56.49	59.23
Consensus taxonomy, earliest year	54.31	79.70	29.17	31.15	52.52	61.79

5. Educational intensity and sectoral performance

The introductory review has highlighted various sources of demand for skilled labor that suggest different associations with capital intensity, demand growth, and labor productivity at the sectoral level. As a final piece of cluster validation, we therefore investigate how the new taxonomy discriminates in terms of these variables. The data stem from the OECD STAN database, which provides a sectoral disaggregation of value added, employment, gross fixed capital formation and labor compensation for a total of 24 countries. The time span is 1992 to 2000.

Without having established an explicit model of sectoral performance, the results presented in this final section can only be tentative, primarily pointing at some promising questions for future research. For the sake of better illustration, Table 6 presents only the coefficients (plus t-value) of a series of ANOVA regressions on various performance measures, which included independent year and country effects in addition to the sector types. All the explanatory variables are categorical and R^2 is generally low. For most variables the regressions can only explain between 5% to 21% of total variation. This should not come as a surprise, since our aim is only to test the discriminatory power of the new taxonomy but not to deliver a comprehensive model of sectoral growth, productivity or investment behaviour. The only exception is the regression on average labor compensation, where the mere categorical sector type, country and time effects explain 62% of the variation. Because of concerns about the likely presence of heteroscedasticity, we also ran a series of non-parametric tests, which generally confirmed the sign and significance of the coefficients in the table.

With the data at hand, we can address a number of tentative empirical predictions. First, human capital theory explains that people invest in education because of their higher expected earnings. If that is true, we *ceteris paribus* expect a positive statistical association between a sector's educational profile and its average labor compensation per workforce. The data are generally consistent with this hypothesis. Compared to the group of sectors with an intermediate educational profile, all industry types with a lower (higher) educational intensity show the corresponding negative (positive) coefficient. However, the relationship is not linear, suggesting that average wages also depend on additional determinants. While the lowest wages are paid in the industries with a very low educational profile, the labor of industries in the category 'low' appears to earn a higher compensation than those in the category of 'intermediate-low'. Conversely, the sectors with an 'intermediate-to-high' educational profile pay the highest wages, while the two remaining groups of 'high' and 'very high' educational intensity are not significantly different from the comparison group.

Second, similar a priori reasoning suggests a positive association between educational intensity and gross fixed capital formation. In the introductory section, we have argued that the extent of interaction with complementarity physical assets (in particular expensive machinery) tends to raise the demand for skilled labor. However, relative to the comparison group of industries with intermediate educational intensity, the ratio of capital investment to the labor force is significantly lower in all the other sector types. The apparent reason is that schooling credentials are an incomplete measure of labor skills. Our finding thus supports the view that much of the presumed capital-skill complementarity rests on informal skills and on-the-job training. This complementarity of physical

assets with intermediate educational degrees (and informal skills) can also explain the previous pattern of labor compensation, where we have found that average wages are particularly high among industries with an intermediate (or intermediate-high) educational intensity.

Third, the aforementioned hypothesis of a *skill-biased technological change* leads us to expect a higher level and/or growth of productivity among the groups of industries with ‘very high’ or ‘high’ educational intensity. But again, the simple conjecture is rejected by the data. Instead, the coefficients for the levels of labor productivity correspond precisely with the pattern observed for capital intensity, while we can hardly discern any significant and interpretable differences for labor productivity growth. One explanation is that much of skill-biased technological change may also happen within industries, irrespective of their initial educational profile. Computers, for example, are a kind of pervasive technological change that is likely to affect the demand for educated labor in a broad range of sectors (and not necessarily those with the highest educational intensity). Part of the effect would then be captured by the independent time effects. A second explanation is the lack of sectoral data on total factor productivity, which certainly would be the more appropriate variable to test for.

Our fourth conjecture is best characterised as a process of *education-biased structural change*, specifically predicting that demand and output grow faster the more education intensive a sector type is. In the introductory section we have presented two likely explanations: (i) the presumed rising income elasticity of demand for education (knowledge) intensive goods and services; and (ii) the increasing integration of world trade, which among the developed OECD countries (with a relative abundance of educated labor) implies a growing specialisation on knowledge intensive sectors. This final conjecture is strongly confirmed by the regression coefficients. Especially the sectors with a very high educational intensity experienced by far the best growth in both variables. Also most industries with a high or intermediate profile did fairly well, while those with a low or very low educational intensity performed worst.¹²

¹² This finding is consistent with Peneder et al. (2003), whose input-output analysis demonstrates a pronounced process of structural change in favor of knowledge intensive sectors, specifically interpreted as ‘quaternarisation’ (in contrast to ‘tertiarisation’, which regards the increasing share of the services sector more generally).

Table 6: Coefficients of sector types in OLS regressions on measures of sectoral performance (incl. fixed sector type, time and country effects)

Industry type	Levels of		Growth of			
	Labor compensation	Capital intensity	Labor productivity	Labor productivity	Value added	Employment
Very high	0.801 (1.31)	-10.224** (11.88)	-73.590** (13.68)	-0.021* (2.37)	0.034** (4.86)	0.047** (11.95)
High	0.040 (0.07)	-9.027** (11.12)	-57.321** (12.49)	-0.008 (1.12)	0.018** (2.83)	0.026** (7.32)
Interm.-high	2.938** (8.41)	-5.946** (12.50)	-49.280** (16.46)	0.005 (0.99)	0.019** (4.26)	0.011** (4.97)
Interm.-low	-8.581** (19.68)	-10.732** (18.60)	-74.724** (19.92)	-0.017** (2.71)	0.010* (1.71)	0.020** (7.01)
Low	-3.397** (9.31)	-10.508** (21.28)	-58.986** (19.12)	-0.001 (0.12)	-0.004 (0.80)	-0.001 (0.33)
Very low	-14.579** (40.79)	-12.543** (26.01)	-83.339** (26.39)	-0.016** (3.07)	-0.032** (6.98)	-0.014** (6.25)
Observations	7221	5022	6720	6720	9855	7252
R ²	0.62	0.21	0.17	0.05	0.09	0.12

Absolute value of t statistics in parentheses; * significant at 5%; ** significant at 1%; The class "intermediate" educational intensity is the comparison group; Average wage rate = total labor compensation / employment; Capital intensity = gross fixed capital formation / employment.

6. Summary and conclusion

The paper develops a new taxonomy of industries based on their educational workforce composition. It provides an empirical tool, that allows the researcher to add analytic structure to micro-level studies of firm behavior, as well as aggregate international and comparative studies – applicable whenever sector dependent characteristics of educational intensity, or its according productive capabilities in the human resource base, are thought to be of importance.

Statistical cluster analysis is applied to data for the USA, Germany, France, the UK and Austria, identifying up to 49 industries that largely correspond to the ISIC 2-digit sectoral breakdown. For practical purposes, the new taxonomy offers various options. First, by virtue of multivariate analysis, we produced a differentiated breakdown of up to seven categories, that range from 'very low' to 'very high' educational intensity. If required, these can also be reintegrated into a smaller number of larger groups (e.g. 'low', 'medium', 'high'). Second, the lack of comparability between different educational systems forced the creation of five separate national taxonomies, which we then synthesized into one common consensus classification. For the purpose of national studies in any of the five countries, presumably applying the taxonomy as conditioning variable in the econometric analysis of firm-level data, the national classification is more appropriate. The 'consensus' classification is the adequate choice for analytic as well as comparative international studies, which presumably are of a more aggregate nature and frequently lack a sufficient number of comparable sectoral data on educational intensity.

The subsequent validation of the taxonomy reveals considerable robustness to variations over time and between countries. Finally, we test for its discriminatory power with respect to various measures of sector performance. While no straightforward pattern could be discerned for capital intensity or the level and growth of labor productivity, we find a clear ranking of value added and employment growth that associates positively with the degree of a sector's educational intensity. The observed differential sectoral growth rates suggest a significant tendency towards education-biased structural change.

Acknowledgements

This paper has benefitted from numerous conversations, helpful comments and invaluable assistance with the national data. Jim Baron, Martin Falk, Edward Lazear, Mary O'Mahony, Laurence Nayman, Bart van Ark, Michela Vecchi, and Thomas Zwick deserve especial thanks. Naturally, none of them bears any responsibility for eventual errors, omissions, or remaining obscurities.

Appendix A: Measures of (dis)similarity

A simple numerical example can demonstrate the differences between the four (dis)similarity functions that we applied in the analysis. Table A.1 provides the values for five hypothetical objects *I* through *V* for the three variables *A*, *B* and *C*. Table A.2 reports the calculated (dis)similarity for four different measures. Figure A.1 offers an additional geometric visualization of the two-dimensional case, in which we only consider the variables *A* and *B*. Objects are characterized in brackets according to their respective co-ordinates. The straight line between two cases corresponds to the Euclidean distance, whereas the city block distance equals the length of the connecting horizontal and vertical lines. The two rays that go from the origin to the respective cases determine the angular separation measure.

Table A.1: A numerical example

Objects	Numerical values of variable		
	<i>A</i>	<i>B</i>	<i>C</i>
<i>I</i>	4.0	2.0	1.0
<i>II</i>	2.0	4.0	3.0
<i>III</i>	3.0	5.0	3.0
<i>IV</i>	3.0	6.0	4.5
<i>V</i>	6.0	4.0	3.0

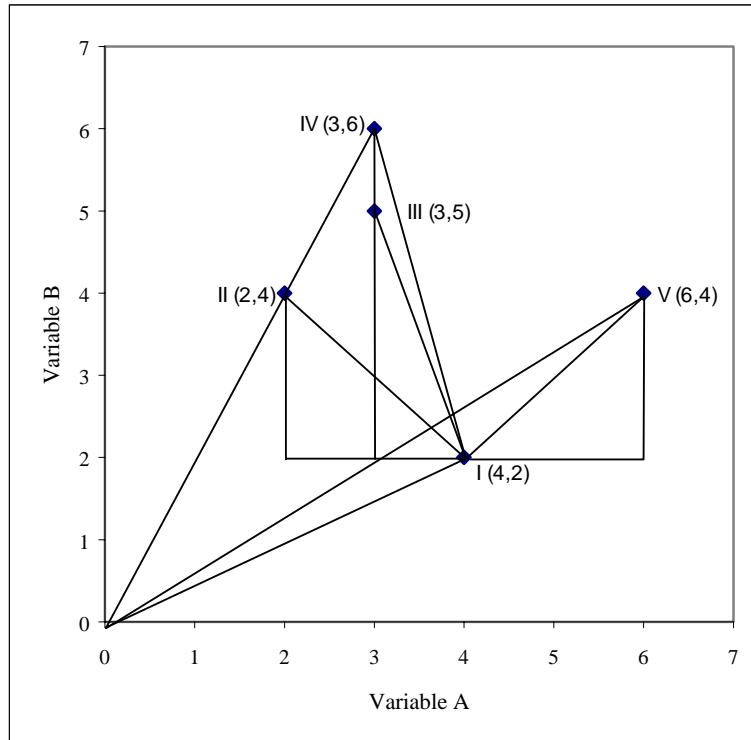
Table A.2: Comparing measures of (dis)similarity of the numerical example

Measure	Comparison of (dis)similarity between object <i>I</i> and ..			
	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>
<i>Euclidean</i>	3.46	3.74	5.41	3.46
<i>City block</i>	6.00	6.00	8.50	6.00
<i>Angular</i>	0.77	0.83	0.77	0.98
<i>Correlation</i>	-0.65	-0.19	-0.65	1.00

The first interesting observation is that the city block measure treats objects *II* and *III* as equally distant from *I*, whereas the Euclidean measure regards the latter as more distant. The simple reason is that we move from a quadratic to a rectangular shape. In contrast, both the angular separation and the correlation coefficient say that relative to *I*, *III* is more similar than *II*. Secondly, for both the Euclidean and the city block distance, case *IV* is more dissimilar to *I* than is case *III* or case *II*. However, when we apply angular separation or the correlation coefficient, *IV* is just as similar to *I* as is *II*, since both locate on the same ray from the origin. Finally, case *V* is an extreme example of the differences between size- and shape-oriented measures. Whereas *I* and *V* are clearly distant in the sense of Euclidean or city block measures (mirroring the distance between *I* and *II*), the two cases are highly similar in the measure of angular separation and even identical, if we apply the correlation coefficient. The reason is that for case *V*, we only add a constant of two units to each of the variables. Since the correlation coefficient is insensitive to mere size displacements, both cases are treated as identical.

Unfortunately, there is no general guideline, which establishes the priority of one measure over another. One might choose the Euclidean distance as the most 'natural' function, but this is only because we are accustomed to imagining objects in Euclidean space. Kaufmann and Rousseeuw (1990), for example, recommend the city block distance instead, because it is not sensitive to outliers. In contrast to both, angular separation and the correlation coefficient are more appropriate when we are interested in similarities in the shape of objects, rather than in the absolute size of the differences. As a practical consideration, a cluster of objects with a similar profile of attributes is often easier to interpret.

Figure A.1: A geometric illustration of differences in (dis)similarity measures



Note: The straight lines between two objects determine the Euclidean distance, the connected horizontal and vertical lines the city block distance and the two rays from the origin the angular separation measure.

Appendix B: Cluster identification for the US national taxonomy

As an instructive example, Figure B.1 presents the hierarchical cluster tree for the USA, using the average linkage method in combination with each of the aforementioned (dis)similarity measures. As is frequently the case with statistical cluster techniques, the most stable patterns appear at the extreme ends of the distribution. Clusters 10 and 11 show the highest degree of educational intensity, with shares of both university and associate degrees far above the average. They are consistently grouped together or located as immediate neighbors in all twelve dendrograms. But the graphical representations also reveal that they are still very distinct among themselves. Cluster 11, which is comprised of sectors such as education, research and development, or computer services, by far exceeds the other sectors in terms of university degrees held by members of its workforce. We consequently label this cluster ‘industries with *very high* educational intensity’. In comparison, cluster 10 comprises industries with a lower but still *high* educational intensity, including for example, financial services, insurance, or the group of other business services.

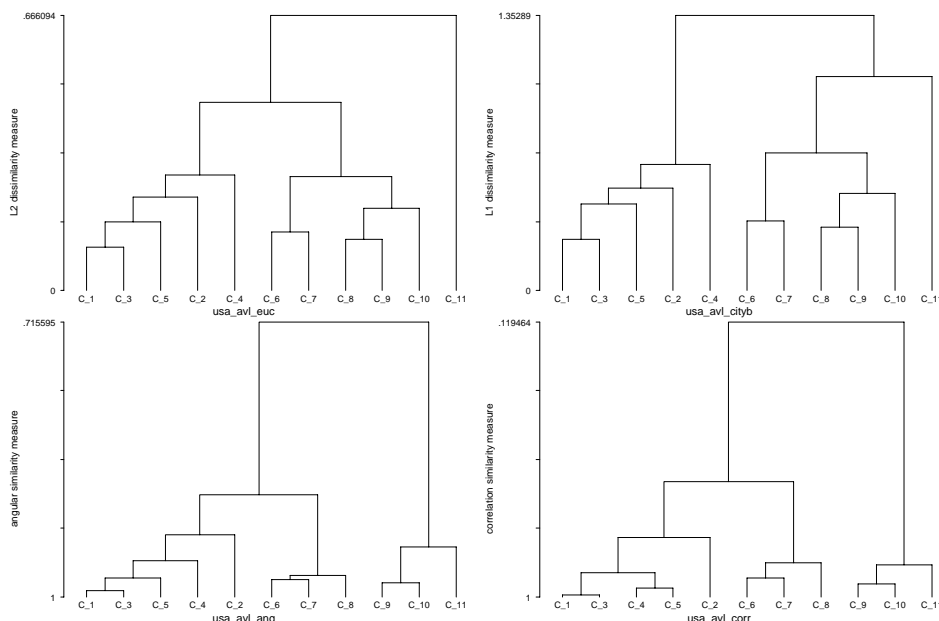
Clusters 6 and 7 exhibit an extremely robust association and are grouped together in all twelve dendrograms. The lack of any pronounced deviations from the average educational composition is the most characteristic observation with respect to those industries with an *intermediate* educational intensity. Examples are mechanical engineering, motor vehicles or retailing.

Clusters 8 and 9 exhibit an above average educational profile, but with more emphasis on associate degrees and fewer university graduates. They are most closely associated in six of the twelve charts, although cluster 9 also appears repeatedly with cluster 10 and cluster 8, together with the intermediate group of clusters 6 and 7. We therefore attach the label ‘*intermediate-to-high* educational intensity’ to this category. Among others, it comprises chemicals, instrument engineering, electrical machinery, and air transport.

The remaining clusters 1 to 5 all exhibit below average educational intensity. Cluster 2 is the most consistent outlier; its workforce has a very low educational profile. Clusters 4 and 5 have the least clear patterns, including among others, food, drink and tobacco, rubber and plastics, basic and fabricated metals, and the group of other inland and water transport. These industries are only consistently close to each other when the correlation coefficient or angular separation are used as similarity measures. The same applies to the city block measure in combination with the complete linkage method. In all other graphs, they are disjoint. Nevertheless, the two are close in 6 of 12

dendrograms. Since the other graphs do not suggest a convincing alternative pattern, we classify them together as industries with *intermediate-to-low* educational intensity.

Figure B.1: USA – Hierarchical Clustering, Average Linkage Method



Clusters 1 and 3 appear together in all twelve cluster trees. They encompass industries such as construction and vehicle maintenance and repair. Their educational profile is distinctly *low*, characterized by above average shares of labor without any formal qualifications, some college (but no degree) or only high school graduation. Finally, cluster 2, which consists of agriculture, forestry and fishing, textiles and clothing, as well as hotels and catering, appears most often in an outlying position. It displays a *very low* educational profile, wherein the share of labor without any formal degrees by far exceeds the other groups.

Appendix C: Supplementary tables

Table C.1. Sectoral evolution by type of educational intensity, USA 1979 - 2000

<i>ISIC</i>	<i>Industry label</i>	1979/80	1981/82	1983/84	1985/86	1987/88	1989/90	1991/92	1993/94	1995/96	1997/98	1999/2000
01-05	Agriculture, forestry, fishing	7	7	7	7	7	7	7	7	7	7	7
10-14	Mining and quarrying	7	5	4	4	4	4	4	4	4	4	4
15-16	Food, drink & tobacco	7	7	7	6	6	6	6	6	6	5	5
17-19	<i>Textiles, leather, footwear & clothing</i>	7	7	7	7	7	7	7	7	7	7	7
21-22	<i>Pulp, paper products; printing, publishing</i>	5	5	5	5	4	4	4	4	3	3	3
23	Mineral oil refining, coke & nuclear fuel	4	4	4	4	4	3	3	3	3	3	3
24	Chemicals	4	4	4	3	3	3	3	3	3	2	3
25	Rubber & plastics	7	6	6	6	6	6	5	5	5	5	5
27-28	<i>Basic metals & fabricated metal products</i>	7	6	6	6	6	6	6	5	5	5	5
29	Mechanical engineering	6	6	5	5	5	5	5	5	5	5	4
30	Computers, office machinery	3	3	2	3	2	2	2	2	1	1	1
31	Electrical machinery & apparatus, nec	5	5	5	4	4	4	4	3	3	3	3
32	Audiovisual apparatus	4	4	4	3	3	3	3	3	3	2	2
33	Instrument engineering	4	4	4	3	4	3	3	3	3	3	3
34	Motor vehicles	6	6	6	6	5	5	5	5	5	5	4
35	Other transport equipment	5	4	4	3	3	3	3	3	3	3	3
36-37	Furniture, miscellaneous manuf.; recycling	7	7	7	7	6	6	6	6	6	6	5
40-41	<i>Electricity, gas & water supply</i>	5	5	5	5	5	4	4	4	4	4	3
45	Construction	7	7	6	6	6	6	6	6	6	6	6
50	Sale & repair of motor vehicles; retail of fuel	7	6	6	6	6	6	6	6	6	6	6
51	Wholesale trade and commission trade	4	4	4	4	4	3	3	3	3	3	3
52	Retail trade; repair (exc. 50)	6	6	5	5	5	5	5	5	5	5	4
55	Hotels & catering	7	7	6	6	6	6	7	7	7	7	7
601	Railways & other inland transport	6	6	6	6	6	5	5	5	5	5	4
602-661	Other inland- and water transport	7	6	6	6	6	6	6	6	6	6	5
62	Air transport	4	4	4	3	3	3	3	3	3	3	3
63	Auxiliary transport activities; travel agencies	4	3	3	3	3	3	3	3	3	3	3
641	Post and courier activities	5	5	5	5	5	5	5	5	5	4	4
642	Telecommunications	5	5	5	4	4	4	4	3	3	3	3
65p67	Financial intermediation (except 66)	4	3	3	3	3	3	3	2	2	2	2
66	Insurance and pension funding	3	3	3	3	3	3	3	3	2	2	2
70	Real estate activities	4	4	3	3	3	3	3	3	3	3	3
71p74	Other business activities	2	2	2	2	2	2	2	2	2	2	2
72	Computer and related activities	2	2	2	2	1	1	1	1	1	1	1
73	Research & development			2	1	1	1	1	1	1	1	1
75	Public admin., defence; social security	3	3	3	3	3	3	3	3	2	2	2
80	Education	1	1	1	1	1	1	1	1	1	1	1
85	Health, social work	4	3	3	3	3	3	3	3	3	3	3
90-99	<i>Other Services</i>	7	4	4	4	4	4	4	4	3	3	3

NB: 1 = very high; 2 = high; 3 = med-high; 4 = intermediate; 5 = med-low; 6 = low; 7 = very low.

Table C.2. OLS regression on employment shares by attainment levels

		University degree	Other formal education	No formal education
Constant		0.118 (12.84)**	0.555 (35.39)**	0.326 (22.13)**
Country				
	UKD	-0.048 (3.89)**	0.061 (2.89)**	-0.013 (0.64)
	FRA	-0.072 (4.93)**	0.005 (0.19)	0.068 (2.87)**
	GER	-0.086 (6.72)**	0.047 (2.16)*	0.039 (1.91)
	AUT	-0.098 (7.47)**	-0.016 (0.70)	0.114 (5.42)**
	USA	c.g.	c.g.	c.g.
Time				
	ca. 1990	-0.040 (9.17)**	-0.038 (5.01)**	0.078 (11.05)**
	ca. 1992	-0.034 (7.80)**	-0.028 (3.68)**	0.062 (8.79)**
	ca. 1994	-0.024 (4.55)**	-0.015 (1.63)	0.039 (4.58)**
	ca. 1996	-0.019 (4.61)**	-0.010 (1.42)	0.028 (4.39)**
	ca. 1998	-0.012 (2.93)**	0.001 (0.09)	0.011 (1.74)
	ca. 2000	c.g.	c.g.	c.g.
Industry type				
	Very high	0.422 (34.02)**	-0.087 (4.13)**	-0.335 (16.88)**
	High	0.285 (22.98)**	0.027 (1.27)	-0.312 (15.71)**
	Interm./High	0.212 (21.13)**	0.085 (5.02)**	-0.297 (18.55)**
	Intermediate	0.123 (12.45)**	0.119 (7.05)**	-0.242 (15.28)**
	Interm./Low	0.030 (2.58)*	0.158 (7.99)**	-0.188 (10.10)**
	Low	0.009 (0.78)	0.099 (5.27)**	-0.108 (6.09)**
	Very Low	c.g.	c.g.	c.g.
Country* Ind. type				
	UKD*Very High	-0.030 (1.72)	-0.016 (0.53)	0.046 (1.64)
	UKD*High	-0.085 (4.86)**	0.048 (1.62)	0.037 (1.32)
	UKD*Interm./High	-0.096 (6.79)**	0.038 (1.57)	0.058 (2.57)*
	UKD*Intermediate	-0.063 (4.52)**	-0.018 (0.75)	0.081 (3.62)**

UKD*Interm./Low	-0.015 (0.90)	-0.053 (1.91)	0.068 (2.59)**
UKD*Low	-0.002 (0.15)	0.008 (0.30)	-0.006 (0.23)
UKD*Very Low	c.g.	c.g.	c.g.
FRA*Very High	-0.096 (4.87)**	0.087 (2.60)**	0.009 (0.28)
FRA*High	-0.099 (5.02)**	0.089 (2.67)**	0.010 (0.30)
FRA*Interm./High	-0.105 (6.34)**	0.034 (1.19)	0.071 (2.69)**
FRA*Intermediate	-0.058 (3.59)**	-0.039 (1.41)	0.097 (3.74)**
FRA*Interm./Low	-0.009 (0.43)	-0.072 (2.10)*	0.081 (2.50)*
FRA*Low	-0.002 (0.08)	-0.044 (1.46)	0.046 (1.60)
FRA*Very Low	c.g.	c.g.	c.g.
GER*Very High	-0.208 (8.59)**	0.113 (2.73)**	0.096 (2.47)*
GER*High	-0.211 (11.17)**	0.173 (5.39)**	0.038 (1.25)
GER*Interm./High	-0.157 (10.54)**	0.085 (3.34)**	0.072 (3.04)**
GER*Intermediate	-0.099 (6.66)**	0.065 (2.59)**	0.033 (1.41)
GER*Interm./Low	-0.024 (1.45)	-0.031 (1.10)	0.055 (2.07)*
GER*Low	-0.002 (0.15)	0.004 (0.15)	-0.002 (0.07)
GER*Very Low	c.g.	c.g.	c.g.
AUT*Very High	-0.097 (5.08)**	0.155 (4.80)**	-0.059 (1.93)
AUT*High	-0.164 (8.21)**	0.199 (5.84)**	-0.034 (1.07)
AUT*Interm./High	-0.129 (8.35)**	0.137 (5.22)**	-0.008 (0.33)
AUT*Intermediate	-0.091 (6.03)**	0.106 (4.14)**	-0.015 (0.62)
AUT*Interm./Low	-0.018 (1.00)	0.047 (1.55)	-0.029 (1.03)
AUT*Low	-0.003 (0.19)	0.089 (3.11)**	-0.085 (3.18)**
AUT*Very Low	c.g.	c.g.	c.g.
Observations	942	942	942
R-squared	0.91	0.60	0.75

Note: Absolute value of t statistics in parentheses; * significant at 5%; ** significant at 1% ; c.g. = comparison group.

References

- Acemoglu, D. (1998), 'Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality', *Quarterly Journal of Economics*, Vol. 113, pp. 1055-1089.
- Acemoglu, D. (2002), 'Technical Change, Inequality, and The Labor Market', *Journal of Economic Literature*, Vol. 40, pp. 7-72.
- Acemoglu, D. and Zilibotti, F. (2001), 'Productivity Differences', *Quarterly Journal of Economics*, Vol. 116, pp. 563-606.
- Akerlof G.A. and Kranton, R.E. (2000), 'Economics and Identity', *Quarterly Journal of Economics*, Vol. 115, pp. 715-753.
- Akerlof G.A. and Kranton, R.E. (2002), 'Identity and Schooling: Some Lessons for the Economics of Education', *Journal of Economic Literature*, Vol. 40, pp. 1167-1201.
- Anderberg, M.R. (1973), *Cluster Analysis for Applications*, Academic Press, New York.
- Arrow, K.J. (1973), 'Higher Education as a Filter', *Journal of Public Economics*, Vol. 2, pp. 193-216.
- Autor, D.H., Katz, L.F. and Krueger, (1998), A.B., 'Computing Inequality: Have Computers Changed the Labor Market?', *Quarterly Journal of Economics*, Vol. 113, pp. 1169-1213.
- Baron, J.N. and Kreps, D.M. (1999), *Strategic Human Resources. Frameworks for General Managers*, Wiley & Sons, New York.
- Becker, G.S. (1964/1975), *Human Capital. A Theoretical and Empirical Analysis, with Special Reference to Education*, NBER and Columbia University Press, New York.
- Berman, E., Bound, J. and Machin, S. (1998), 'Implications of Skill-Biased Technological Change: International Evidence', *Quarterly Journal of Economics*, Vol. 113, pp. 1245-1279.
- Bresnahan, T.F., Brynjolfsson, E. and Hitt, L.M. (2002), 'Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence', *Quarterly Journal of Economics*, Vol. 117, pp. 339-376.
- Brewer, D.J., Eide, E.R. and Ehrenberg, R.G. (1999), 'Does It Pay to Attend an Elite Private College? Cross-Cohort Evidence on the Effects of College Type on Earnings', *The Journal of Human Resources*, Vol. 34, pp. 104-123.
- Card, D. (1999), 'The Causal Effect of Education on Earnings', in: Ashenfelter, O., Card, D. (eds.), *Handbook of Labor Economics*, Vol. III, Elsevier, Amsterdam, pp. 1801-1862.
- Card, D. and Krueger, A.B. (1992), 'Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States', *Journal of Political Economy*, Vol. 100, pp. 1-40
- Caroli, E. and Van Reenen, J. (2001), 'Skill-Biased Organizational Change? Evidence from a Panel of British and French Establishments', *Quarterly Journal of Economics*, Vol. 116, pp. 1449-1492.
- Caselli, F. (1999), 'Technological Revolutions', *American Economic Review*, Vol. 89, pp. 78-102.
- Chun, H. (2003), 'Information Technology and the Demand for Educated Workers: Disentangling the Impacts of Adoption versus Use', *The Review of Economics and Statistics*, Vol. 85, pp. 1-8.
- Dearden L., Reed, H., and Van Reenen, J. (2000), 'Who Gains When Workers Train? Training and Corporate Productivity in a Panel of British Industries', IFS Working Paper, 00/04, London.
- de la Fuente, A. and Doménech, R. (2002), 'Human Capital in Growth Regressions: How Much Difference Does Data Quality Make? An Update and Further Results', mimeo.

- Falk, M., Seim, K. (2001), 'The Impact of Information Technology on High-Skilled Labor in Services: Evidence from Firm-Level Panel Data', *Economics of Innovation and New Technology*, Vol. 10, pp. 289-324.
- Forbes, K.J. (2001), 'Skill classification does matter: estimating the relationship between trade flows and wage inequality', *Journal of International Trade & Economic Development*, Vol. 10, pp. 175-209.
- Freeman, R.B. (1986), 'Demand for Education', in: Ashenfelter, O., Layard, R., *Handbook of Labor Economics*, Vol. I, Elsevier, Amsterdam, pp. 357-386.
- Gordon, A.D. (1999), *Classification*, 2nd ed., Chapman & Hall, Boca Raton.
- Hamermesh, D.S. (1993), *Labor Demand*, Princeton University Press, Princeton.
- Haskel, J.E. and Slaughter, M.J. (2002), 'Does the Sector Bias of Skill-biased Technical Change Explain Changing Skill Premia?', *European Economic Review*, Vol. 46, pp. 1757-1783.
- Kahn, J.A. and Lim J.-S. (1998), 'Skilled Labor-Augmenting Technical Progress in U.S. Manufacturing', *The Quarterly Journal of Economics*, Vol. 113, pp. 1281-1308.
- Kaufmann, L. and Rousseeuw, P.J. (1990), *Finding Groups in Data. An Introduction to Cluster Analysis*, Wiley, New York.
- Kerckhoff, A.C. and Dylan, M. (1999), 'Problems with International Measures of Education', *The Journal of Socio-Economics*, Vol. 28, pp. 759 - 775.
- Krueger, A. (1993), 'How Computers Have Changed the Wage Structure: Evidence from Micro data, 1984-1989', *Quarterly Journal of Economics*, Vol. 108, pp. 33-60.
- Lazear, E.P. (1995), *Personnel Economics*, MIT Press, Cambridge MA.
- Lazear, E.P. (1998), *Personnel Economics for Managers*, Wiley & Sons, New York.
- Machin, S. and Van Reenen, J. (1998), 'Technology and Changes in Skill Structure: Evidence from Seven OECD Countries', *Quarterly Journal of Economics*, Vol. 113, pp. 1215-1244.
- Peneder, M. (2001), *Entrepreneurial Competition and Industrial Location*, Edward Elgar, Cheltenham, UK.
- Peneder, M. (2003), 'Industry Classifications. Aim, Scope and Techniques', *Journal of Industry, Competition and Trade*, Vol. 3, pp. 109-129.
- Peneder, M., Kaniovski, S. and Dachs, B. (2003), 'What Follows Tertiarisation? Structural Change and the Role of Knowledge-Based Services', *Service Industries Journal*, Vol. 23, pp. 47-66.
- Romesburg, H.C. (1984), *Cluster Analysis for Researchers*, Waldsworth Inc., Belmont.
- Schultz, Th.W. (1960), 'Capital Formation by Education', *Journal of Political Economy*, Vol. 68, pp. 571-583.
- Schultz, Th.W. (1961), 'Investment in Human Capital: Reply', *American Economic Review*, Vol. 51, pp. 1035-1039.
- Sianesi, B. and Van Reenen, J. (2002), 'The Returns to Education: A Review of the Empirical Macro-Economic Literature', IFS Working Paper 02/05, London.
- Spence, M. (1973), 'Job Market Signalling', *Quarterly Journal of Economics*, Vol. 87, pp. 355-374.
- Spence, M. (2002), 'Signalling in Retrospect and the Informational Structure of Markets', *American Economic Review*, Vol. 92, pp. 434-459.
- Wolff, E.N. (2003), 'Skills and Changing Comparative Advantage', *The Review of Economics and Statistics*, Vol. 85, pp. 77-93.