Working Paper No. 159

IT skills in vocational training curricula and labour market outcomes

Fabienne Kiener, Ann-Sophie Gnehm, Simon Clematide and Uschi Backes-Gellner

Universität Zürich
IBW – Institut für Betriebswirtschaftslehre

UNIVERSITÄT BERN

Working Paper No. 159

IT skills in vocational training curricula and labour market outcomes

Fabienne Kiener, Ann-Sophie Gnehm, Simon Clematide and Uschi Backes-Gellner

September 2022 (first version: February 2019)

This paper was previously circulated under the title "Different Types of IT Skills in Occupational Training Curricula and Labor Market Outcomes" (2019).

Published as: “IT skills in vocational training curricula and labour market outcomes.“ *Journal of Education and Work,* 35(2022)6-7: 614-640. By Fabienne Kiener, Ann-Sophie Gnehm, Simon Clematide and Uschi Backes-Gellner.

DOI: https://doi.org/10.1080/13639080.2022.2126968

The Swiss Leading House on Economics of Education, Firm Behavior and Training Policies is a Research Program of the Swiss State Secretariat for Education, Research, and Innovation (SERI).

www.economics-of-education.ch

IT skills in vocational training curricula and labour market outcomes[1]

Fabienne Kiener[2], Ann-Sophie Gnehm[3], Simon Clematide[4]
and Uschi Backes-Gellner[5]

September 2022

Abstract

We use vocational training curricula to investigate how IT skills are trained within broader *skills packages* and how these relate to labour market outcomes. Skills packages are the typical combinations of IT skills (e.g., CNC) and technical or nontechnical skills (e.g., material sciences or work safety) that are jointly required in the real world and occur in training curricula. This broadened perspective of teaching IT skills offers new insights into how digital skills can be successfully integrated into future education and training programs. We use legally binding vocational education and training (VET) curricula of dual apprenticeship training in Switzerland. We apply natural language processing methods to analyse the extensive curriculum texts, which meticulously define the skills that have to be taught. We identify four typical skills packages, each of which are centred around one of four different types of IT skill (CNC/CAD, control technologies, system technologies, IT-applications). Our empirical analyses show that VET graduates trained in these skills packages receive positive labour market outcomes compared to VET graduates without these skills packages. Moreover, we find that the positive outcomes are not just driven by differences in cognitive skill requirements of the respective occupations.

Keywords: IT skills, information technologies, apprenticeship, training, curricula; **JEL Classification:** I26, J24, O33

---

[1] This study was partly funded by the Swiss State Secretariat for Education, Research, and Innovation (SERI) through its "Leading House VPET-ECON: A Research Center on the Economics of Education, Firm Behavior and Training Policies". This paper has benefitted from many audiences and exchanges. In particular, we would like to thank the following individuals for their help and support: Simone Balestra, Dietmar Harhoff, Simon Janssen, Edward Lazear, Jens Mohrenweiser, Samuel Muehlemann, Harald Pfeifer, Paul Ryan, Guido Schwerdt, conference participants in New York City (SASE) and Augsburg (COPE), seminar participants at the University of Zurich for helpful comments, Natalie Reid for language consulting, and the Swiss Federal Statistical Office for data provision. The purchased data on cognitive requirement levels of each VET occupation (Anforderungsprofile) belongs to the Swiss Trade Association.

[2] University of Zurich, Department of Business Administration.
[3] University of Zurich, Institute of Sociology
[4] University of Zurich, Institute of Computational Linguistics
[5] Corresponding author. University of Zurich, Department of Business Administration. Email: backes-gellner@business.uzh.ch

1. Introduction

Given the increasing use of information technologies (IT) in the labour market, IT training is growing in importance. However, valuable IT training should constitute more than just learning a particular IT skill without any context, because on-the-job IT skills are typically combined with a set of related skills that complement the IT technology and knowledge. For example, when workers learn computer numerical control (CNC), they would also need to have good knowledge of lathing and milling, material properties, how to read manufacturing plans or how to execute production plans with CNC to be able to use the technology effectively. The question of teaching IT skills combined with a package of complementary skills has so far been neglected in the economics literature. However, knowing how such skills packages look like is crucial to design training programs that are intended to provide workers with skills required during times of increased digitalisation (see e.g., Spöttl and Windelband, 2021, focusing on how nine occupational fields will need to adapt due to the fourth industrial revolution).

So far, empirical studies on IT skills and labour market outcomes focus on just IT software, programming languages, or hardware in job postings (Atalay et al. 2018; Buchmann, Buchs, and Gnehm 2020), on elaborate self-reports of IT skills (Borghans and ter Weel 2006), or on IT use for basic problem-solving (Falck, Heimisch-Roecker, and Wiederhold 2020). However, an empirical analysis of how effective IT training for the workplace is and should be combined with other technical or nontechnical skills, i.e., skills packages with IT skills, is still missing in the literature.

In this paper, we investigate how IT skills are taught in broader skills packages in extensive vocational education and training (VET) training curricula in the context of three- or four-year dual apprenticeship training programs in Switzerland, which cover approximately two-thirds of the Swiss labour force, i.e., all middle-skilled workers. We analyse how such skills packages with IT skills are related to labour market outcomes of graduates of the respective training programs. Such dual VET at the upper-secondary education level is based on legally binding curricula for on-the-job training in a company that employs the apprentice (approx. 80% of the time) and school-based learning in a special vocational school (approx. 20% of the time). Finishing such a dual VET degree offers well-defined pathways to vocational tertiary education options (see Backes-Gellner and

Geel 2014 for the Swiss system, and Jaik and Wolter 2019 for the occupational choice process in Switzerland).

Our skills data include text as data from 166 vocational training curricula with an average of 44 pages each. Studies that manually analyse curricula from dual VET have shown in other contexts that their content, for example its degree of specificity, is important for labour market outcomes (e.g., Eggenberger, Rinawi, and Backes-Gellner 2018; Eggenberger and Backes-Gellner 2020).[6] We dig deeper and look at the content and how particular skills are combined in typical 'skills packages with IT skills' and how these skill packages relate to labour market outcomes of the respective VET graduates. Our labour market data are from the Swiss Social Protection and Labour Market Survey (SESAM), which contains individual data from a representative sample of the Swiss working population.

As the curricula of the 166 Swiss occupations are very detailed and thus provide extensive text bodies (in total 8,102 pages), we apply natural language processing (NLP) methods, i.e., machine learning applied to texts, to analyse the content of these curriculum texts. In doing so, we built on a recent strand of empirical literature in economics, sociology and other related disciplines that has applied automatic text analysis—primarily searching for specific keywords—to job postings (Buchmann, Buchs, and Gnehm 2020; Brown and Souto-Otero 2020; Deming and Kahn 2017; Grinis 2019; Schultheiss et al. 2018). We extend this literature by introducing a new type of text, i.e., vocational training curricula, and by applying an innovative methodological approach for our content analysis, i.e., topic modelling.[7] It uses an *unsupervised* NLP algorithm that can be applied to a large body of text without predefined vocabulary, thereby helping to recognize text patterns in vast curriculum texts and to identify typical skills packages. Our unsupervised algorithm clearly detects four different skills packages with IT, each of them comprising different IT skills and other technical and nontechnical skills: first, a skills package with

[6] Eggenberger and Backes-Gellner (2020) analyse the interaction of specificity and IT skills, focusing on IT skill as single skills (see subsection 'Concept of skills packages with IT' for more details).

[7] Topic modelling recognizes patterns, called topics, in vast quantities of texts through an algorithm that performs a soft clustering of documents. Because we use curricula, our topics represent skills packages. The most important hyper-parameter that topic modelling requires is the number of topics. We evaluate the optimal number of topics with a method that tests how coherent the topics are in terms of the similarity or differences among words. In our case, twelve topics are optimal. We then use an interpretation tool to analyse which topics contain IT terms. Four of our twelve topics represent skills package with IT, thereby topic modeling results in measures for each skills package with IT in each curriculum (see section 'Identifying "skills packages with IT' in vocational training curricula' for more details).

CNC/Computer-Aided Design (CAD) including e.g., planning of material, manufacturing, testing; second, a skills package with system technologies including e.g., process orientation, assembly, technical documentation; third, a skills package with IT-applications including e.g., project management, independent handling, orders; and fourth, a skills package with control technologies including e.g., knowledge on moulding, cutting tools, practice-oriented understanding.

Our empirical analyses show how the four identified skills packages with IT are associated with labour market outcomes, such as wages and employment.[8] Our results show that workers trained in a skills package with IT have higher wages and a higher employment probability, compared to those without training in such a skills packages with IT. In the wage analysis, when comparing the size effects of the four different skills packages with each other, the results indicate that the skills package with IT-applications has higher wage returns than the three other skills packages with IT, i.e., with control technologies, with CNC/CAD, and with system technologies. While the importance of IT-applications is also shown in previous studies (e.g., Atalay et al. 2018 demonstrate the increase of skills in IT-applications required in job postings), our results also emphasize that the skills package with IT-applications comprises not only IT-applications but also project management.

Additional analyses address the potential concern that skills packages with IT skills may only occur in cognitively demanding occupations, in which case we would measure only the labour market effect of cognitively demanding occupations and not capture the effect of IT skills packages. To proxy the cognitive skill requirements of an occupation (independently from the skills packages from the text analysis), we use an external database that provides the 'cognitive requirement levels' for each Swiss VET occupation (Goetze and Aksu 2018). Our additional empirical analyses reveal that our previous results do not change structurally, i.e., the positive relationships between our skills packages with IT

[8] One important assumption of our empirical analyses is that the extent to which skills are mentioned in occupational training curricula corresponds to the importance of the skills during the training, i.e., a curriculum, which e.g., specifies the training of CNC skills in more detail and across more pages than other skills, provides a more intensive training in these CNC skills than a curriculum, which specifies the training of CNC skills with comparatively less detail and across fewer pages. This seems reasonable given the institutional context of VET in Switzerland as described in section 'Approaches to measure skills.' This assumption is also used in previous research (e.g., Eggenberger, Rinawi, and Backes-Gellner 2018) and has shown to provide valid results for labour market analyses.

and their labour market outcomes are not just the effect of cognitively more demanding occupations.

In sum, our paper provides evidence that efficient IT training is taught in skills packages that combine IT skills with a range of other, complementary skills. Conceptually, IT skills packages contain the respective IT skill with complementary skills such as technical, organizational, social or other noncognitive skills. These additional skills make the use of the IT skills more efficient due to complementarity effects. For example, learning CNC is only valuable if workers have knowledge on material properties, are able to read manufacturing plans and are familiar with testing. Workers having the combination of such skills achieve a higher quality in the production process or are better able to efficiently address complications in the production process. In contrast, workers with only CNC skills and limited or no knowledge of material properties—or who are not able to read manufacturing plans, or who are not familiar with testing—would fail to do so. Thus workers who possess the whole skills package with the respective IT skills are more valuable for firms than workers who only possess a single IT skill without the corresponding complementary skills. As a result, we expect firms to pay higher wages for workers possessing such complementary skills packages. Empirically, we find four different skills packages with IT. We show what skills they consist of and demonstrate how they are related to labour market outcomes. Thus, our study provides novel conceptual ideas and empirical evidence on the importance of teaching coherent packages of skills that include important IT skills together with the most relevant complementary technical and nontechnical skills.

2. Background on skills packages with IT

In this section, we propose our concept of skills packages with IT. Subsequently, we provide an overview of two approaches using text data sources, curricula and job postings, to measure skills.

2.1 Concept of skills packages with IT

We outline the concept of skills packages with IT in three steps. First, we present the empirical research on IT skills, i.e., all skills that help workers handle, apply, and use information technologies (Falck, Heimisch-Roecker, and Wiederhold 2020). Second, we

add a training perspective that considers skills packages with IT. Third, we conceptually link the empirical studies and the training perspective.

Empirical studies on IT skills and labour market outcomes go back to Krueger (1993) who studied how wages correlate with the use of computers in the workplace. In contrast, DiNardo and Pischke (1997) discussed that the approach of measuring workers' computer use does not adequately represent IT skills. Recent studies contribute additional approaches to measuring IT skills. This research—for example—analyses the effects of mentioning IT software, languages, or hardware in job postings (Atalay et al. 2018; Buchmann, Buchs, and Gnehm 2020) or how IT workers value using the newest technologies drawing on employer reviews (Tambe, Ye, and Cappelli 2020). Further studies use workers' self-reported IT skills (Borghans and ter Weel 2006) or assess basic problem-solving using IT (Falck, Heimisch-Roecker, and Wiederhold 2020). To date, empirical research tends to view IT skills as handling one or a few particular IT skills and misses to consider how IT skills are actually taught (i.e., a training perspective).

Shedding light on IT skills from a training perspective, we propose that IT skills are taught with other skills during IT training and that different forms of these IT trainings exist. Vocational training curricula reflect IT training and skills.[9] Curricula, such as those of Swiss VET occupations define the content of the training, i.e., they contain detailed descriptions of skills. Extensive final examinations guarantee that when entering the labour market, VET graduates possess the skills required for their occupation (Eggenberger, Rinawi, and Backes-Gellner 2018). Such curricula show how IT skills are trained effectively, thereby demonstrating the training perspective of IT skills. To illustrate the training perspective, we describe another example—besides the one of CNC skills mentioned in the introduction: Learning how to use software is not only about the functionality of the software but also on how to accomplish specific professional tasks (e.g., a commercial employee who does the bookkeeping) with this software in the context of the production or service process—meaning that a specific IT task is always connected to other tasks and knowledge. Only with the combination of these skills, workers are able to efficiently perform the bookkeeping for a department or firm. Thus, simply learning the functionality

[9] While a study by Spöttl and Windelband (2021) discusses how the fourth industrial revolution requires nine occupational fields to adapt in the future, we focus on how IT training is currently implemented in VET curricula.

of a software does not suffice, but a worker needs to combine software skills with, for example, basic knowledge of accounting.

To conceptually link the empirical studies and the training perspective, we extend the theoretical arguments drawn from Lazear's (2009) skills-weights approach. This approach introduces the idea that training programs or workers' human capital should be considered as 'skill bundles' that contain a variety of different single skills with different weights. For example, Eggenberger, Rinawi, and Backes-Gellner (2018) analyse skill bundles at the occupational level; i.e., an occupational training program consists of a combination of single skills that are needed in the respective occupation. The study of Eggenberger, Rinawi, and Backes-Gellner (2018) was recently extended by a new study introducing IT skills as a particular category of single skills (Eggenberger and Backes-Gellner 2020). They distinguish between generic and expert IT skills and show that *generic* IT skills positively correlate with earnings after involuntary separations for workers with specific occupational skill bundles. However, this study does not investigate how IT skills are combined with other, technical or nontechnical skills in different occupations. Our approach argues that instead of just considering IT skills as single skills, it is—particularly from a training perspective—also important to consider that single IT skills are not or should not be taught individually but together with a set of other skills that are used complementary in the typical workplaces and occupations for such IT skills. In our approach, each occupation has a large number of single skills that are typically combined in empirically clearly identifiable skills packages. Some of these skills packages include IT, some not, some skills packages have a higher IT content and others a lower IT content. We thus have an empirical profile of the skills packages for each occupation. For each occupation, we have a weight for all its skills packages with IT. This weight represents the importance of this IT skills package in the training of this occupation and in this sense also provides a training perspective on IT skills.

2.2 Approaches to measure skills

For our empirical analyses, we decided to use training curricula as a text data source instead of job postings, which have mainly been used in previous literature. We do so because analysing training curricula has two major benefits over analysing job postings for our purpose: First, while the texts of job postings mainly *name* a couple of skills that

are currently required at the workplace for a certain occupation, the texts of VET training curricula (in the countries with dual VET programs such as Switzerland or Germany) extensively *define* and *elaborate* the skills that are taught and required in an occupation (and that workers are guaranteed to have because of mandatory examinations).[10] Second, in contrast to job postings that can always only name a few, currently important skills due to the very limited space of job ads, the curriculum texts describe all skills that workers acquire during their extensive training period (three to four years). Because of their goal and brevity, job postings can always only just mention the skills that are the most important for the recruiting process at the particular point in time, but these skills are of course surrounded by many other skills that remain unnamed. The unnamed skills may still be implied in the job ads because they are part of common knowledge in a particular industry or occupational field—but for a text analysis they would nevertheless be invisible. Therefore, we prefer to use the more comprehensive VET curriculum texts for our analysis of skills packages rather than job postings.

To operationalize and empirically measure skills from texts using machine learning, we also discuss different types of methods that can be used to extract the information from the texts in job postings (e.g., Atalay et al. 2018; Buchmann, Buchs, and Gnehm 2020; Brown and Souto, 2020; Deming and Kahn 2017) or in curriculum texts (e.g., Eggenberger, Rinawi, and Backes-Gellner 2018; Eggenberger and Backes-Gellner 2020; Janssen and Mohrenweiser 2018).

The first approach focuses on measuring the skill requirements in job postings, ads that occur naturally as data in text form, with machine-learning methods. These methods process unstructured texts to organize the information inherent in them (Gentzkow, Kelly, and Taddy 2019), because compared to conventional data sources (e.g., occupational

[10] As mentioned in the introduction section of this paper, a crucial assumption of our analyses is that if a curriculum extensively (i.e., across more pages and in more detail) describes a particular skill, this skill is trained more intensively (i.e., more training time is invested in this skill). In contrast, another skill, which is only briefly (i.e., across few pages and in less detailed) described in the curriculum, this skill is trained less extensively (i.e., only little training time is invested in this skill). This seems reasonable because the curricula are legally binding documents that need to be followed by all training firms; thus, to meet the legally required standards, firms will closely follow the content that is prescribed in the training curricula. Furthermore, we assume that workers who graduated from the respective occupations indeed acquired these skills because they need to pass a final assessment (with written, oral and practical assessment parts), which assesses whether the students acquired all the skills that are prescribed in the curricula. Therefore, we are confident that the curriculum texts and the shares of different skills therein correspond to the extent of skills that workers acquire. Moreover, previous research has also taken this assumption and has delivered crucial findings on skills and labour market returns (e.g., Eggenberger, Rinawi, and Backes-Gellner 2018, Eggenberger and Backes-Gellner 2020).

skills databases), original texts as data are likely to contain more detailed information than, for example, survey information. Atalay et al. (2020), Buchmann, Buchs, and Gnehm (2020), Brown and Souto-Otero (2020), Deming and Kahn (2017), Grinis (2019), and Schultheiss et al. (2018) show that job posting texts contain relevant labour market information. Atalay et al. (2020) even demonstrate that job posting texts reflect the evolution of work better than the growth or decline of job titles.

A second approach is based on vocational training curricula that define the content of training and thus offer insights into the acquired skills. For example, this method is applied for studying the dual VET systems in Germany and Switzerland (Eggenberger, Rinawi, and Backes-Gellner 2018; Jansen, Grip, and Kriechel 2017; Janssen and Mohrenweiser 2018; Eggenberger and Backes-Gellner 2020). In previous studies, researchers manually analyse the skills described in vocational training curricula and then link the skills with labour market outcomes (Eggenberger, Rinawi, and Backes-Gellner 2018). In contrast to these manual analyses, our study is the first to apply machine-learning methods to curricula, and following our study, recently, some studies have also analysed curricula using machine-learning methods (Eggenberger & Backes-Gellner, 2020, or Kiener, Gnehm, and Backes-Gellner 2020).

Both strands of literature, the studies on job postings (e.g., Atalay et al. 2018; Buchmann, Buchs, and Gnehm 2020; Deming and Kahn 2017) and the studies on curricula (e.g., Eggenberger, Rinawi, and Backes-Gellner 2018; Eggenberger and Backes-Gellner 2020; Janssen and Mohrenweiser 2018), have in common that they define their skills a priori, i.e., they used predefined vocabulary and—if they use NLP methods and not a manual analysis—they primarily apply supervised NLP methods. Using predefined vocabulary is useful when the research focuses on one or a few particular skills they are interested in. In this case, researchers can use specific keywords that capture the single skills instead of a primarily data-driven unsupervised NLP method like ours (see section ‘Identifying skills packages with IT in training curricula’). Given that our paper is interested in finding patterns of skill combinations in curricula that we cannot define a priori but want to identify empirically, i.e., skills packages that are typically taught together in the context of vocational training, we decided not to use a method that requires a priori definitions of a predefined (small selection) of single skills but rather a method that avoids the need for a priori definitions. Thus, we use a topic modelling method that does not require a priori

definitions or categorization but is able to detect hidden, i.e., latent, within-text structures called topics (details are provided in the section 'Identifying skills packages with IT in training curricula').

3. Data

The data we use to measure skills packages with IT, hereafter called 'skills data', are vocational training curricula, i.e., Swiss VET curricula. We use the curriculum texts for VET training as published on the online directory of the Swiss confederation for legally mandated dual vocational education and training programs on the Federal level.[11] In total, we cover the most important 166 Swiss VET curricula. The curricula are not only extensive (a total of 8,102 pages, averaging 44 pages for each curriculum) but also rich in content (a total of approximately 1.5 million words, averaging 8,030 words for each curriculum).

In addition to the skills data, we use SESAM data on the labour market. The data set comprises detailed information for a representative sample of the Swiss population. The SESAM links data from the Swiss Labour Force Survey (SLFS) with information from different social insurance registers and has a rolling panel structure (Federal Statistical Office 2011). The Federal Statistical Office connects the information from the survey with administrative data. We build a panel between 2010 and 2016 similar to Balestra and Backes-Gellner (2017), Eggenberger, Rinawi, and Backes-Gellner (2018) or Eggenberger and Backes-Gellner (2020), who constructed a panel for earlier years. In 2010, the Statistics Office changed the survey frequency thus we only use the data after 2010 to have a consistent time series for our analysis.[12]

Among the variables in the SESAM are our dependent variables, i.e., the administrative wage data on the gross income from employment per year and the employment status. The SESAM data also include characteristics such as age, gender, worktime percentage, marital status, and nationality. Relevant variables for the matching and sampling

[11] https://www.becc.admin.ch/becc/public/bvz/beruf/grundbildungen (downloaded in summer 2018). The download comprised curricula (in German, 'Bildungsplan') of the VET occupations with a Federal Diploma (in German, 'Eidgenössisches Fähigkeitszeugnis EFZ'). Some VET occupations within the same occupational field have one curriculum, for example the orchardist (in German, 'Obstgärtner/in') and the farmer (in German, 'Landwirt/in').

[12] Since 2010, the SLFS has interviewed each individual five times in one-and-a-half years; before they interviewed them only once a year but for five consecutive years.

procedure are two variables: the 'level of education' variable (we use with VET diploma at the upper-secondary level) and the variable on the 'training occupation' (i.e., in which occupation the worker was originally trained).

We match the skills data and the individual labour market data at the occupational level.[13] The labour market data from the SESAM includes the variable on the original training occupation, which for individuals with a VET education corresponds to the occupations covered by the curricula. We link the codes for the occupations from the two datasets following Eggenberger, Rinawi, and Backes-Gellner (2018), who did a similar matching for earlier curricula.[14]

Having described the data, we present how we obtain the measures for different skills packages with IT from our skills data using the NLP method in the following section. Subsequently, we present the labour market analyses.

4. Identifying 'skills packages with IT' in vocational training curricula

We use NLP methods that combine techniques from machine learning with linguistics and help analyse vast quantities of text to identify the skills packages in curriculum texts. Such computational methods for text analyses create many new opportunities for economic research (Gentzkow, Kelly, and Taddy 2019). Previous text analyses based on manual coding needed a priori definitions and categories; in other words, the categories were defined before an individual coded the texts (Quinn et al. 2010). One example of such manual categorization of skills is found in the study of Eggenberger, Rinawi, and Backes-Gellner (2018) who manually categorized Swiss occupational skill bundles. The topic modelling approach we use,[15] does not require such a priori definitions or categorization but detects hidden, i.e., latent, within-text structures called topics (Grimmer and Stewart 2013). We use these topics to represent the skills packages that are prescribed and defined in the VET curricula. We are the first who measure skills packages as an

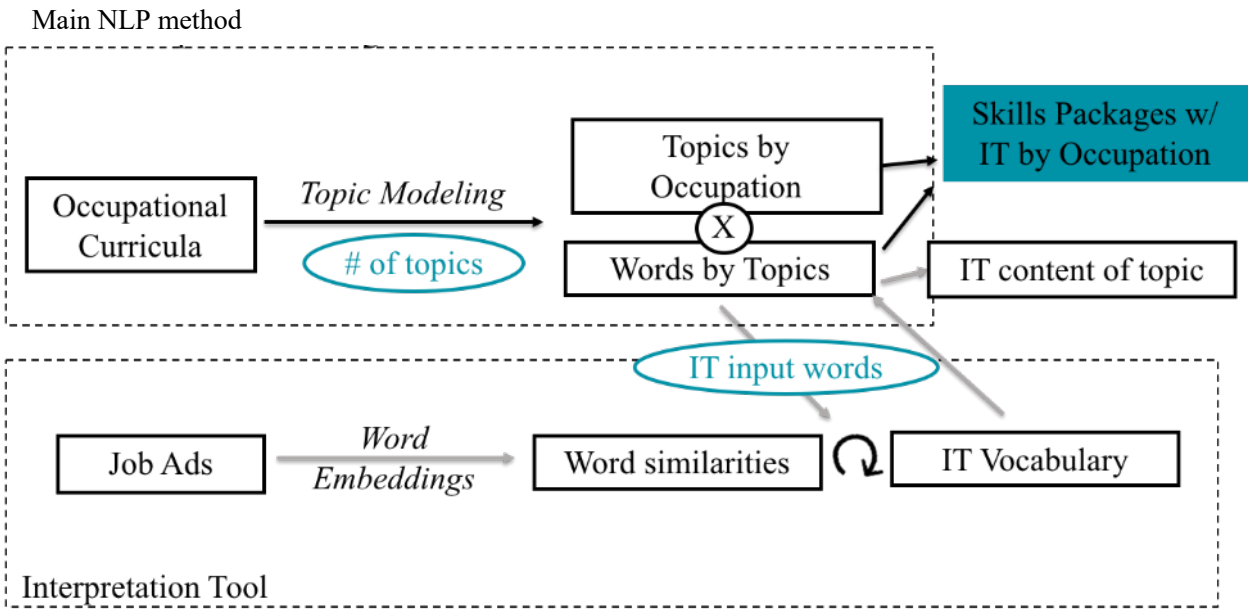
[13] We follow the assumption from Eggenberger, Rinawi, and Backes-Gellner (2018): VET workers with graduation previous to the observation period update their skills according to the recent curriculum.
[14] We match 134 of the VET occupations from the skills data to 215 original training occupations in the labour market data.
[15] The computations were carried out in cooperation with the Computational Linguistics Department at the University of Zurich.

operationalization of the theoretical concept of skills packages as outlined in the section 'Concept of skills packages with IT'.

Topic modelling has never before been applied to vocational training curricula. Thus far, applications of topic modelling have been used to analyse Twitter, news data, speeches, or scientific abstracts in political, communication and marketing research (Amado et al. 2018; Benites-Lazaro, Giatti, and Giarolla 2018; Maier et al. 2018; Reisenbichler and Reutterer 2019). An example of a previous application focusing on economics research studies how topics in economic research evolve over time through applying topic modelling to full texts of articles in the JSTOR database (Ambrosino et al. 2018).

In our application, the topics represent skills packages because we use curriculum texts as a database. The following schematic illustration presents our NLP process. The process starts with occupational curricula and ends in a measure for skills packages with IT by occupation. We apply two methods: first, topic modelling, i.e., the main NLP method (explained in subsection 'Main NLP method'), and second, the interpretation tool (explained in subsection 'Interpretation tool'). Finally, the last subsection shows the resulting skills packages with IT.

Figure 1. NLP methods.



Source: Authors' schematic illustration of the use of NLP methods. The italics show the algorithms and techniques. The oval text boxes represent hyper-parameters and input by the researchers.

4.1 Main NLP method

Our main NLP method is topic modelling, which we apply to vocational training curricula. The topic-modelling algorithm detects topics (patterns) within vast quantities of text. We use a topic-modelling algorithm called nonnegative matrix factorization (NMF) (for a mathematical explanation, see Gillis 2014 or for applications, see Lee and Seung 1999; Shahnaz et al. 2006; Tjioe et al. 2008; Quinn et al. 2010; Yan et al. 2013). NMF uses matrix factorization, i.e., a method in linear algebra. Applying NMF to curricula factorizes the curricula into two matrices: 'Words by Topics', which shows the words importance for each topic, and 'Topics by Documents', which shows the distribution of the topics over the documents, i.e., in our case curricula. As a simplified example, words by topics shows words on 'developing something', and topics by documents shows that the topic occurs in the training curriculum of an information technologist or in the curriculum of a photographer (see appendix A for a more elaborate example).

To calculate the matrices 'Words by Topics' and 'Topics by Documents', the topic-modelling algorithm requires the number of topics as the most important hyper-parameter set by the researcher. We evaluate the optimal number of topics with a method that tests how coherent the topics are in terms of the similarity or differences among words. Specifically, we use a word-intrusion method where humans detect random 'intrusion words' in words by topic. The better the detection of the intrusion word, the more semantically coherent the topic is (Chang et al. 2009). Typically, researchers choose three possible numbers of topics for testing (e.g., Maier et al. 2018). We chose the possible numbers of topics according to previous research,[16] and executed the word intrusion procedure with 12, 24, and 48 topics. In our case, 12 topics were an optimal number based on the semantic coherence of the topics. Up to this point, the decisions are rather generic and the IT content of a topic has not been of any relevance. Only in the next step we analyse the content of a topic and evaluate whether they contain IT content or not. To do so, we use a so-called interpretation tool.

[16] Based on previous research and suggestions by Simon Clematide from the computational linguistics institute at the University of Zurich, we use 12, 24 and 48 topics as a start for our curriculum analysis.

4.2 Interpretation tool

The interpretation tool helps identify the content of the skills packages, i.e., topics, with IT or with other content. Because we want to find IT vocabulary in topics, our interpretation tool is similar to a dictionary. To automatically build a dictionary, we want to draw from a large text database that is not the curriculum texts themselves but that contains texts belonging to a domain similar to curricula. We decided to use job postings due to availability and closeness to the labour market. The interpretation tool starts with a corpus of 1.15 million online job ads from 2014 through 2018 (i.e., the task description, required skills and personal characteristics).[17] We use dense continuous word representations, 'word embeddings', as a technique to calculate word similarities. This technique relies on the context of words, and words in similar contexts have a similar meaning.[18]

The interpretation tool helps us analyse whether the 100 most important words in each topic match the IT vocabulary of the job ads. To do that, we take some words from the matrix 'Words by Topics' that are related to IT. We name them 'IT input words.' Then, the interpretation tool shows words similar to the IT input words, resulting in an IT vocabulary that we apply to the words by topics (see the complete list of IT vocabulary in the appendix A).[19]

Overall, the interpretation tool supports us in defining the content of the topics through an automatically created distributional dictionary. Thus the interpretation tool helps us make our contribution to unsupervised topic modelling, which is—according to Gentzkow, Kelly, and Taddy (2019) and Quinn et al. (2010)—the interpretation of the topics. Our resulting skills packages do not exclusively depend on subjective interpretations of researchers but also on the objective dictionary from job ads. By doing so, we extend the traditional use of topic modelling in social sciences and the analysis of skills in curriculum texts.

[17] The corpus comes from the Swiss Job Market Monitor, which is associated with the institute of sociology at the University of Zurich.
[18] In particular, word embeddings represent each word in a vector space. They statistically capture the meaning of words by their use in the text, i.e., reflect semantic properties of words. Based on the vector space, we calculate similarities between words by measuring their cosine distance.
[19] We use word2vec by Mikolov et al.(2013) applying a minimum cosine similarity of 0.6.

4.3 The resulting skills packages with IT

Our NLP methods result in a measure of skills packages with IT that can be found in the occupational curricula. Conceptually, we argued that skills packages with IT contain IT skills and other, complementary technical and nontechnical skills and our empirical measures of skills packages with IT follow this concept. Empirically, after the algorithm calculates the skills packages and assigns a weight for each skills package, i.e., topic, for each curriculum (topics over documents), the interpretation tool helps to identify the skills packages with IT (analysing the words by topic). In total, our method detects four skills packages with IT (for skills packages without IT, see table B1 in the appendix). Within these four skills packages with IT, we also have the words that occur in addition to the IT vocabulary, i.e., further skills and competences in this skills package that are not related to IT. The table below shows the words, which describe IT skills or other skills, in the skills packages with IT as well as the most important occupations that contain the respective skills package.

Table 1. Skills packages with IT.

Skills package with IT	*Important words*	*Occupations with largest weight of the skills package with IT (numeric weight)*
Skills Package w/ CNC/CAD	3 IT skills: CNC, manufacturing technologies, CAD Other important skills and competences in this skills package: manufacturing, documenting, production documentation, planning material, testing, work safety, checking quality, fulfil order, comply with environmental protection, produce…	Mechanical engineer (0.59) Machine Operator & Automation Technician (0.54) Precision Optician (0.54) Automation Engineer (0.54)
Skills Package w/ system technologies	10 IT skills: electrical system technologies, installation, system technologies, communication system, system, building automation, apprentices install Other important skills and competences in this skills package: electrotechnical, technical documentation, assembly, process orientation, learning strategy, work technique, circuit, self-responsibility, energy distribution, coaxial system…	Electrical Installer (0.66) Electrical Planner (0.64) Assembly Electrician (0.58) Telematics Technician (0.55)
Skills Package w/ IT-applications	45 IT skills: ICT, implement, application, server, user, development of applications, configure, server service, ICT System Operation, network infrastructure, install, network, system technologies, website, software, … Other important skills and competences in this skills package: handle small project, order, project, independent handling, solutions, explain independently, user handling, self-reflection, customers, working economically…	Information Technologist (0.59) ICT Expert (0.45) Multimedia electronics technician (0.09) Mediamatic (0.09)
Skills Package w/ control technologies	5 IT skills: CAM, machine technologies, manufacturing technologies machine technologies control technologies, machine technologies Other important skills and competences in this skills package: permanent mould, lost mould, cutting tool, moulding material, professionally according to specifications, practice-oriented understanding, set up a machine, centrifugal jet, adhere to solidification time…	Casting Technologist (0.69) Casting Moulder (0.68) Cutler (0.54) Mould Builder (0.53)

Notes: Authors' calculations from the curricula database using the NLP methods. Important words stem from the matrix words by topics. Top occupations and their weights from the matrix topics over documents.

We name the first package in Table 1 ‘Skills Package with CNC/CAD’ because the following three IT words (in the 100 most important words of the skills package) occur: CNC, manufacturing technologies and CAD. Other important words in the skills package are manufacturing, documenting, production documentation, and planning material. Thus the skills package with CNC/CAD typically also includes complementary skills, such as planning the material. A mechanical engineer has the highest weight for this skills package. The weight here represents the importance of the skills package for the occupation and is determined by the algorithm. The weight has a value between zero and one. The skills package with CNC/CAD occurs in many occupations: Sixty out of 166 occupations have a positive weight for this skills package.

We name the second package ‘Skills package with System Technologies’ because it mainly includes IT words about system technologies. Other important words are electrotechnical, technical documentation, assembly, and process orientation. The skills package with system technologies is important for electrical occupations. However, more than 40 out of 166 occupations have a positive weight for this skills package, i.e., it is to a smaller extent rather common in the Swiss labour market.

We name the third package ‘Skills package with IT-Applications’ because it has very high content of words about applications of IT. This skills package also comprises skills in e.g., project management. This skills package has a high weight in the curriculum of—and is thus very important for—information technologists. In other occupations, this skills package is less important: For example, the occupation with the fourth highest weight for the skills package with IT-applications, mediamatic (an occupation that combines marketing and information technologies), already has a much lower weight than the information technologist. On the other hand, this skills package with IT-applications also occurs in many occupations but with a rather low weight (54 out of 166 occupations).

We name the fourth package ‘Skills package with Control Technologies’ because it includes a variety of IT words about control technologies. The non-IT words in this package are mainly about casting and moulding. Fifty-one out of 166 occupations have a positive weight for the skills package with control technologies so it is also rather broadly used on the Swiss labour market.

In brief, the detailed illustration of the skills packages with IT shows that each skills package contains different IT and different other technical and nontechnical skills. In

preparation of the labour market analyses, we further emphasize that our skills packages with IT differ from each other (because skills packages with IT are not significantly correlated). Simultaneously, each skills package with IT occurs in a substantial share of the occupations (between 43 and 60 out of 166 occupations).

While the main contribution of identifying our four skills packages with IT is to show that—and how—IT skills are trained in context, we can also relate them to results from previous literature on IT skills and compare how their definition of IT skills correspond to our skills packages. We take as an example three studies that investigate IT skills, which are closest to the ones we also identified. Although these studies investigate only single IT skills, they provide a first indication of the importance of *our skills packages* because *their single IT skills* are also comprised in our skills packages:

First, Atalay et al. (2018) study the increased mentioning of office software or programming languages in job postings. Atalay et al.'s (2018) findings show that firms demand these single IT skills more and more between 1960 to 2000. In our paper, we identify the skills package with IT-applications, which comprises the single IT skills of Atalay et al.'s (2018) study as well as other skills such as project management or the processing of orders. Second, Buchmann, Buchs, and Gnehm (2020) investigate, among others, CAD and CNC skills as industry-specific single IT skills, documenting that job postings increasingly mention these single skills in the last three decades. In our paper, we detect a skills package with CNC/CAD, which includes these single IT skills as well as other skills such as how to plan material, manufacturing skills or testing. Third, Eggenberger and Backes-Gellner (2020) present, among others, an IT skill they call 'developing microcontroller systems', showing how this expert IT skill relates to job separations of workers with specific skill bundles. This single IT skill also occurs in our skills package with system technologies, which includes process orientation, assembly and technical documentation.

Overall, these examples demonstrate that previous research focusing on single IT skills identifies important IT skills that are also part of our skills packages with IT. However, our skills packages with IT go beyond the concept of single skills and show with which other (technical or nontechnical) skills these IT skills are typically combined and thus can be considered as complementary skills. In the next section, we show that the combination of these skills is indeed valuable on the labour market.

5. Skills packages with IT and labour market outcomes

We investigate the relationship between the four different skills packages with IT and labour market outcomes, such as wages and employment probability. For each of the two labour market outcomes, we start with the sample, explain the specification, and present the results.

5.1 Wages

Our sample includes the data from 2010 through 2016 for the working population, defined as individuals aged 18 to 64[20] who receive a positive labour income, i.e., who were employed.[21] We study the curricula of VET occupations, and thus we restrict the sample to graduates of vocational education and training.[22] As the dataset includes the worktime percentage of each worker, we can calculate the full-time equivalents of the annual administrative wages. We exclude observations with outlying wages (i.e., first and last percentiles). Moreover, we only include individuals with values for our control variables of age, gender, and nationality.

Our sample comprises 89,259 observations from 56,014 individuals, 57% of whom are men and 43% of whom are women. More than half of the individuals are Swiss citizens. The annual average wage in full-time equivalents over all observations is CHF 80,851 (see figure B1 histogram of wages in the appendix). We standardize the measures of the different skills packages with IT on the sample. Table 2 displays summary statistics.

[20] Typically, the students receive their VET diploma at age 18, and 64 corresponds with the earliest full-pension retirement age.
[21] We do not include observations with zero wages, which the study of Balestra and Backes-Gellner (2017) recommend to include when analyzing earning losses, because the dependent variable are log-wages. However, the results remain the same when we include zero wages (of unemployed individuals), do not calculate full time equivalents and forego the log-function.
[22] Vocational Education and Training Federal Diploma of VET ('Berufslehre EFZ') and Vocational matura ('Berufsmatura')

Table 2. Summary statistics: Wage sample.

Variable	*Mean*	*St. dev.*	*Min*	*Max*	*N*
Annual wages	80,851	34,014	6,313	236,000	89,259
Skills package with CNC/CAD (standardized)	0.00	1	-0.33	4.01	89,259
Skills package with system technologies (standardized)	0.00	1	-0.22	4.96	89,259
Skills package with IT-applications (standardized)	0.00	1	-0.31	6.19	89,259
Skills package with control technologies (standardized)	0.00	1	-0.55	20.03	89,259
Age	43.62	11.53	18	64	89,259
Female	0.43	0.50	0	1	89,259
Swiss	0.69	0.46	0	1	89,259

Notes: Authors' calculations of the summary statistics from the wage sample, based on the their skills measures and the Labour Market Survey (SESAM), 2010–2016. The skills measures are standardized on this sample. The displayed values relate to overall observations.

To analyse the relationship between wages and skills packages with IT, we use a Mincer-type wage regression. The independent variable is the log of annual wages (in full-time equivalents). Measures of the skills packages with IT, i.e., the skills package with CNC/CAD, the skills package with system technologies, the skills package with IT-applications, and the skills package with control technologies, are based on original training occupations of each individual, which should not change over time. We control for the individual characteristics of gender, age, age squared, and being Swiss. We decided to only control for characteristics that are not influenced by the training occupation such that we do not include 'bad controls' (e.g., industry or firm size, see Angrist and Pischke 2009, for the definition of 'bad controls'). We also control for the year of the observation (i.e., year fixed effects).

Our labour market data allow for constructing a rolling panel structure (Balestra and Backes-Gellner 2017; Eggenberger, Rinawi, and Backes-Gellner 2018; Eggenberger and

Backes-Gellner 2020). Because we have observations of individuals for more than one year, we use a pooled ordinary least squares (OLS) model. We cluster standard errors on training occupations, because the 'level of treatment', i.e., skills packages with IT are measured in the curricula of training occupations.

(I) $$\log(wage_{i,t}) = \beta_0 + \beta_{1-4} SkillsPackage\ w/IT_i + \beta_5 gender_i + \beta_6 age_{i,t} + \beta_7 age_{i,t}^2 + \beta_8 Swiss_i + \beta_{9-16} year_{i,t} + \varepsilon_{i,t}$$

, where $t = 2010, \dots, 2016;\ i = 1, \dots, N.$

We analyse whether the wage return differs between (a) workers with any skills package with IT vs. workers without and (b) workers that have one or the other of the four different skills packages with IT. We first regress the log of wage on each skills package with IT separately and then include all skills packages with IT.

The first main result constitutes that workers with any skills packages with IT receive higher wage returns than workers without skills packages with IT: For each of the four skills package with IT (i.e., the skills package with CNC/CAD, the skills package with system technologies, the skills package with IT-applications, and the skills package with control technologies), we find statistically significant positive relationships with wages in comparison to workers without any IT skills package (see table 3).

The second main result shows that the positive wage returns of workers with any skills package with IT differ depending on which skills package with IT they have: Comparing the size of the coefficients, the fifth regression in table 3 displays a coefficient of 0.058 for the skills package with IT-applications (i.e., workers with the skills package with IT-applications receive higher wages by 5.8% compared to workers without an IT skills package), followed by the coefficient of 0.054 for the skills package with control technologies, the coefficient of 0.027 for the skills package with CNC/CAD, and finally, the coefficient of 0.019 for the skills package with system technologies. To further explore whether the wage returns (i.e., coefficients in fifth regression in table 3) between the different skills packages with IT indeed differ from each other, we also investigate their pairwise F-tests: The coefficients for the different skills package with IT are significantly

different from each other (see table B2 in the appendix).[23] In other words, each skills package with IT receives a different wage return.

Having shown the main results, we also investigate whether the relationships between the skills packages with IT and wages depend on the controls. Thus we add our controls one by one to the regression with skills packages with IT (see table B3 in appendix). The analysis shows that the coefficients for the skills packages with IT mostly change when adding gender (and only slightly for age or being Swiss). A concern of our skills packages with IT could be that other skills packages (those without IT) would diminish the previously shown main wage results. However, even when we add dummies for the skills packages without IT to our specification, the coefficients for the skills packages with IT remain positive and significant (see table B4 in appendix). Thus our additional analyses support our main result of statistically significant positive relationships between each skills package with IT and wages.

[23] An exception is the coefficient of the skills package with control technology, which is not statistically different from the skills package with CNC/CAD or the skills package with IT-applications, meaning that the positive wage return of the skills package with control technology is not different from the one with CNC/CAD and the one with IT-applications.

Table 3. Wages and skills packages with IT: Main analysis.

VARIABLES	(1) *log wage*	(2) *log wage*	(3) *log wage*	(4) *log wage*	(5) *log wage*
Skills package w/ CNC/CAD	0.018**				0.027***
	(0.007)				(0.007)
Skills package w/ system technologies		0.007			0.019***
		(0.005)			(0.005)
Skills package w/ IT-applications			0.057***		0.058***
			(0.006)		(0.005)
Skills package w/ control technologies				0.054***	0.054***
				(0.015)	(0.015)
Controls for gender, age, age², Swiss, years	Yes	Yes	Yes	Yes	Yes
Constant	9.645***	9.646***	9.666***	9.670***	9.671***
	(0.057)	(0.059)	(0.056)	(0.058)	(0.060)
Observations	89,259	89,259	89,259	89,259	89,259
Number of individuals	56,014	56,014	56,014	56,014	56,014
R^2 overall	0.118	0.117	0.131	0.128	0.145
R^2 between individuals	0.123	0.122	0.136	0.133	0.150

Notes: Authors' calculations, based on their skills measures and the SESAM, 2010–2016. Robust standard errors in parentheses clustered on training occupation. *** p<0.01, ** p<0.05, * p<0.1. Reading example regression (5) coefficient CNC/CAD: 'An increase of 1 standard deviation of the skills package with CNC/CAD is associated with a 2.7% wage increase.'

5.2 Employment

The employment analysis uses a different sample than the wage analysis, because for analysing the employment probability, we do not only focus on the working population but we can also include unemployed individuals.[24] We still restrict the sample to individuals aged 18 to 64 with VET as the highest education level. As in the previous sample, we only include individuals with values for the control variables (age, gender, Swiss nationality).[25]

[24] See footnote 21, which further argues why—for the previous wage analyses—we use a different sample (the working population without zero wages).

[25] For the employment analyses, we do not exclude the first and last percentile of the wage distribution because for these analyses, we have a broad sample that also includes unemployed individuals.

Our new sample comprises 99,822 observations from 61,752 individuals, of whom 92% (8%) are employed (unemployed). A total of 56% are men and 44% are women. Seventy percent of the individuals have a Swiss citizenship. We standardized the measures of the different skills packages with IT on the employment sample. Table 4 displays summary statistics of the variables.

Table 4. Summary statistics: Employment sample.

Variable	*Mean*	*St. dev.*	*Min*	*Max*	*N*
Employment	0.92	0.27	0	1	99,822
Skills package with CNC/CAD (standardized)	0.00	1	-0.32	4.13	99,822
Skills package with system technologies (standardized)	0.00	1	-0.22	5.12	99,822
Skills package with IT-applications (standardized)	0.00	1	-0.31	6.30	99,822
Skills package with control technologies (standardized)	0.00	1	-0.56	20.20	99,822
Age	44	11.49	18	64	99,822
Female	0.44	0.50	0	1	99,822
Swiss	0.70	0.46	0	1	99,822

Notes: Authors' calculations of the summary statistics from the employment sample. Data based on their skills measures and the SESAM, 2010–2016. The skills measures are standardized on this sample. The displayed values relate to overall observations.

We estimate a simple linear probability model to analyse the relationship between employment and skills packages with IT. The independent variable is binary, being employed or not, and we choose the linear probability model following Angrist and Pischke (2009). We use the measures for different skills packages with IT standardized on the employment sample. We also control for age, age squared, gender, Swiss nationality and years. Again, we cluster standard errors on training occupations due to the same reasoning as in the wage analyses (the 'level of treatment'—skills packages with IT—are measured in the curricula of training occupations).

(II) $\Pr(employed = 1|\mathrm{X}) = \beta_0 + \beta_{1-4} SkillsPackagew/IT_i + \beta_5 gender_i + \beta_6 age_{i,t} + \beta_7 age_{i,t}^2 + \beta_8 Swiss_i + \beta_{9-16} year_{i,t} + \varepsilon_{i,t}$

, where $t = 2010, \ldots, 2016;\ i = 1, \ldots, N$

Our analyses show whether the probability of being employed differs between (a) workers with any skills package with IT vs. workers without and (b) workers that have one or the other of the four different skills packages with IT. The first main result is that workers with any skills package with IT have a higher employment probability than workers without a skills package with IT (see table 5).

The second main result shows that between workers with different skills packages with IT, the employment probability varies: The fifth regression in table 5 shows the coefficient of 0.012 for the skills package with CNC/CAD (i.e., workers with the skills package with CNC/CAD have a higher employment probability by 1.2 percentage points compared to workers without an IT skills package), followed by the coefficient of 0.008 for the skills package with system technologies, the coefficient of 0.007 for the skills package with control technologies, and finally, the coefficient of 0.005 for the skills package with IT-applications. In contrast to the results of the wage returns, the skills package with IT-applications correlates with the lowest probability of being employed. To further explore whether the employment probabilities (i.e., coefficients in fifth regression in table 5) between the different skills packages with IT indeed differ from each other, we also investigate their pairwise F-tests: The coefficients of the skills package with CNC/CAD, the one with IT-applications and the one with system technologies are statistically significant different from each other; however, the coefficient of the skills package with control technologies is not different from the others (see table B5 in the appendix).

To investigate whether the relationship between the skills packages with IT and the employment probability depends on the controls, we show in additional regressions that the coefficients of the skills packages with IT change only slightly when adding the controls one by one (see table B6 in the appendix). In parallel to the wage analysis, we also address the possible concern whether including skills packages without IT diminishes the main employment result; thus we add dummies for the skills packages without IT to our specification. The coefficients for the skills packages with IT remain positive and significant, with the exception of the coefficient of the skills package with control technologies,

which is no longer significant possibly due to potential multicollinearity (see table B7 in appendix). Therefore, our additional analyses further demonstrate our main result: Individuals trained in skills packages with IT have a higher probability of being employed.

Table 5. Employment and skills packages with IT: Main analysis.

VARIABLES	(1) *employed*	(2) *employed*	(3) *employed*	(4) *employed*	(5) *employed*
Skills package w/ CNC/CAD	0.010***				0.012***
	(0.003)				(0.003)
Skills package w/ system technologies		0.006***			0.008***
		(0.002)			(0.002)
Skills package w/ IT-applications			0.004**		0.005**
			(0.002)		(0.002)
Skills package w/ control technologies				0.006*	0.007*
				(0.004)	(0.004)
Controls for gender, age, age², Swiss, years	Yes	Yes	Yes	Yes	Yes
Constant	0.931***	0.931***	0.934***	0.936***	0.930***
	(0.020)	(0.020)	(0.021)	(0.020)	(0.019)
Observations	99,822	99,822	99,822	99,822	99,822
Number of individuals	61,752	61,752	61,752	61,752	61,752
R^2 overall	0.0105	0.00968	0.00952	0.00975	0.0125
R^2 between individuals	0.0129	0.0118	0.0117	0.0119	0.0146

Notes: Authors' calculations, based on their skills measures and the SESAM, 2010–2016. Robust standard errors in parentheses clustered on training occupation. *** p<0.01, ** p<0.05, * p<0.1. Reading example regression (5) coefficient skills package with CNC/CAD: 'One standard deviation in the skills package with CNC/CAD is associated with an increase in the probability of being employed by 0.012.'

6. Additional analyses of skills packages with IT, requirement levels, and labour market outcomes

In this section, we address one potential concern about our analyses in more detail, that is, the cognitive level of skills and tasks that are required in certain occupations. The question is whether skills packages with IT only occur in cognitively demanding occupations. If so, then the positive relationships of skills packages with IT and labour market

outcomes may be only due to the higher cognitive ability of the individuals. Besides addressing this concern, our additional analysis adds to the studies that use PIACC data on IT use for problem-solving as IT skills (Falck, Heimisch-Roecker, and Wiederhold 2020) and in another study as a part of cognitive skills (Shields and Sandoval Hernandez 2020). To evaluate this concern, we include the intellectual requirement levels of occupations in our analyses. The requirement level reflects the cognitive difficulty level of a VET occupation resulting from the average of an expert assessment (between 1 and 100) in math, school language, natural sciences, and a first foreign language (see Goetze and Aksu 2018, for the methodology, and Jaik and Wolter 2019 or Wolter and Zumbuehl 2017 for studies using the data).[26]

Because the requirement levels are not available for all VET occupations, we had to build a slightly smaller wage[27] and employment[28] sample for these analyses. Occupations that include IT may be cognitively demanding; thus we expect positive correlations between requirement levels and skills packages with IT. The correlation table below shows that requirement levels are positively correlated with the skills package with CNC/CAD, the skills package with IT-applications, and the skills package with control technologies. In contrast, the skills package with system technologies is slightly negatively correlated with the requirement levels, meaning that occupations with the skills package with system technologies have a low requirement level. Thus not every skills package with IT needs to be positively correlated with requirement levels—a result that further emphasizes the importance of distinguishing different skills packages with IT.

[26] We take the average requirement level across all subjects (math, school language, natural sciences, and first foreign language), because we want to capture as many dimensions of 'cognition' as possible (e.g., more than only mathematical requirement level). However, the empirical results are robust when only including the mathematical requirement level.

[27] We drop 4,865 (5.5 percent) observations with occupations without data on requirement levels for our new wage sample. The dropped observations have a wage mean of 83,953 and standard deviation of 33,429 compared to original wage sample of 80,851 and a standard deviation of 34,014. As for the summary statistics, the requirement levels have a mean of 44 and a standard deviation of 8.4 (minimum 24, maximum 63).

[28] We drop 5,457 (5.5 percent) observations with occupations without data on requirement levels for our new employment sample. The dropped observations have an employment mean of 0.913 and a standard deviation of 0.282 compared to main employment sample with an employment mean 0.92 and a standard deviation of 0.27. As for the summary statistics, the requirement levels have a mean of 44 and a standard deviation of 8.4 (minimum 24, maximum 63).

Table 6. Skills packages with IT and requirement levels: Correlations.

	Skills package w/ CNC/CAD	*Skills package w/ system technologies*	*Skills package w/ IT-applications*	*Skills package w/ control technologies*
Wage sample:				
requirement levels	0.286*	-0.096*	0.300	0.466*
Employment sample:				
requirement levels	0.283*	-0.091*	0.301*	0.473*

Notes: Authors' calculations, based on their skills measures and the SESAM, 2016. Pairwise correlations with a Bonferroni correction. Only data on 2016 included. * means statistical significance on the 0.01 significance level.

We compare our baseline regression of wages with a regression that includes requirement levels. The regression with requirement levels shows that three out of the four skills packages with IT still have positive and significant coefficients. The central coefficient in the wage regression, that for the skills package with IT-applications, is smaller than it was (now 0.035, before 0.059, i.e., adding requirement levels, workers with the skills package with IT-applications receive higher wages by 3.5% compared to workers without any IT skills packages) but is still larger than the coefficients for the other skills packages. Additionally, the analysis supports that our measures of skills packages with IT have additional explanatory power for wages of different occupations even when including their intellectual requirement levels (see table 7, comparing R squared of regressions two and three).

Table 7. Wages, skills packages with IT, and requirement levels: Additional analysis.

VARIABLES	(1) *log wage*	(2) *log wage*	(3) *log wage*
Skills package w/ CNC/CAD	0.029***		-0.003
	(0.007)		(0.011)
Skills package w/ system technologies	0.020***		0.014***
	(0.005)		(0.006)
Skills package w/ IT-applications	0.059***		0.035***
	(0.005)		(0.008)
Skills package w/ control technologies	0.059***		0.025*
	(0.016)		(0.014)
Requirement levels		0.011***	0.009***
		(0.002)	(0.002)
Controls for gender, age, age², Swiss, years	Yes	Yes	Yes
Constant	9.669***	9.182***	9.292***
	(0.063)	(0.080)	(0.099)
Observations	84,394	84,394	84,394
Number of individuals	52,989	52,989	52,989
R^2 overall	0.147	0.153	0.160
R^2 between individuals	0.153	0.159	0.166

Notes: Authors' calculations, based on their skills measures and the SESAM, 2010–2016. Robust standard errors in parentheses clustered on training occupation. *** p<0.01, ** p<0.05, * p<0.1.

Similar to the wage regression, adding requirement levels to the employment regression also adds explanatory power (see table 8, comparing R squared of regressions two and three). The coefficients of 0.007 for the skills package with system technologies is still significantly positive. This coefficient means that—adding requirement levels —workers with the skills package with system technologies have a higher employment probability by 0.7 percentage points compared to workers without any IT skills packages. In contrast to the coefficient of the skills package with system technologies, the coefficients of the other skills packages with IT are nonsignificant when including requirement levels. Comparing the wage regressions and the employment regressions, the requirement levels have more impact on the relationships between skills packages with IT and the employment

probability than the ones between skills packages with IT and wages. In other words, while individuals possessing skills packages with IT have higher wages irrespective of the cognitive requirements, skills packages with IT are not the only driver of a higher employment probability, because cognitive requirements also explain part of the relationship.

Table 8. Employment, skills packages with IT, and requirement levels: Additional analysis.

VARIABLES	(1) *employed*	(2) *employed*	(3) *employed*
Skills package w/ CNC/CAD	0.012***		0.006
	(0.003)		(0.005)
Skills package w/ system technologies	0.008***		0.007***
	(0.002)		(0.002)
Skills package w/ IT-applications	0.005**		0.001
	(0.002)		(0.004)
Skills package w/ control technologies	0.007*		0.000
	(0.004)		(0.004)
Requirement levels		0.002***	0.002
		(0.001)	(0.001)
Controls for gender, age, age², Swiss, years	Yes	Yes	Yes
Constant	0.930***	0.853***	0.859***
	(0.020)	(0.033)	(0.047)
Observations	94,365	94,365	94,365
Number of individuals	58,398	58,398	58,398
R^2 overall	0.0126	0.0133	0.0143
R^2 between individuals	0.0148	0.0155	0.0165

Source: Authors' calculations, based on their skills measures and the SESAM, 2010–2016. Robust standard errors in parentheses clustered on training occupation. *** p<0.01, ** p<0.05, * p<0.1.

Our additional analyses further emphasize the need for a differentiated view of the different skills packages with IT. Additionally, requirement levels do not completely remove

the explanatory power of the skills packages with IT. Thus the additional analyses address the concern that skills packages with IT only occur in cognitively demanding occupations. In other words, these analyses demonstrate that the positive labour market outcomes associated with skills packages with IT are not primarily driven by cognitively demanding occupations.

7. Conclusion

This paper first conceptually introduces 'skills packages' as an important analytical category for labour market outcomes. We argue that labour market-oriented IT training needs to teach IT skills in combination with complementary technical and nontechnical skills. This results in training curricula typically containing comprehensive skills packages with IT skills rather than just single IT skills without the necessary complementary skills. Second, we empirically identify four skills packages with IT skills that are currently relevant in the Swiss labour market, i.e., in a highly innovative and internationally competitive economy. We show the major elements of these skills packages with IT and how they correspond to labour market outcomes of individuals graduating with the respective skills packages. The four skills packages with IT based on applying an unsupervised NLP algorithm to occupational training curricula are the following: first, a skills package that revolves around *CNC/CAD* skills and includes e.g., knowledge of materials, testing, and work safety; second, a skills package that revolves around *system technologies*, including e.g., process orientation or technical documentation; third, a skills package that revolves around *IT-applications*, including e.g., project management, assembling and working economically; fourth, a skills package that revolves around *control technologies*, including e.g., knowledge on moulding or cutting tools. Our empirical findings emphasize that the combination of IT skills with other technical and nontechnical skills is important and indicate how efficient IT training programs can be developed in the future.

Our labour market analyses show higher wage returns and a higher employment probability for workers with any skills package with IT compared to workers without an IT skills package. Moreover, the positive wage returns and employment probabilities vary across the different skills packages with IT. Thus the labour market analyses further stress that a differentiated view is important. Furthermore, we examine how our main results are connected to the cognitive requirement levels of the tasks in a given occupation and

the workers selecting into these occupations, which we proxy with the intellectual requirement levels of each occupation. Our results show that even though the cognitive requirement levels and skills packages with IT are correlated, the different skills packages with IT still improve the explanatory power regarding labour market outcomes. Thus, our results are not only driven by differences in cognitive abilities.

While economic returns are one important—and rather easily obtainable—indicator for the value of such skills packages, they may also bring other types of values, e.g., the individuals' satisfaction of receiving such training or firms' cost-benefit of training such skills packages. These are therefore certainly valuable avenues for future research. However, we argue that economic returns for the individuals are already a very important aspect of possible returns to study the value of skills packages with IT, because they (a) contribute to the worker's financial well-being—possibly across their whole labour market career—and (b) are a strong indicator of how firms value such skills packages. The importance of economic returns is also stressed by further research that studies noncognitive skills and wage returns demonstrating the relevance of investigating such relationships (Kiener, Gnehm, and Backes-Gellner 2020).

Our main contribution to the economics literature is (a) to show that typical IT training consists of skills packages with IT, (b) to illustrate how these skills packages with IT look like, and (c) to emphasize their importance for labour market outcomes such as wages and the employment probability. Whereas the economics literature has dealt with IT skills as 'single' skills (e.g., Atalay et al. 2018, Buchmann, Buchs, and Gnehm 2020, Borghans and ter Weel 2006, Falck, Heimisch-Roecker, and Wiederhold 2020, or Eggenberger and Backes-Gellner 2020), we conceptually and empirically study skills packages with IT, which go beyond the view of 'single' IT skills but rather stress the relevance of a 'package of skills' (i.e., a combination of IT skills with complementary technical and nontechnical skills). Given that we add a novel training perspective by investigating skills packages with IT and do not know in advance how the skills packages with IT look like, we do not have an a priori, specific definition of the skills at the start of the empirical investigations.

Methodologically, this paper also introduces a novel way of measuring skills packages with IT showing how skills can be measured in vast quantities of texts with an unsupervised NLP method. Using a large body of text as a data source is a recent development in economics or sociological research and our method is applicable to text as data (but not

to other data sources). We particularly contribute to the current research on skills in job postings because we apply a novel method that needs fewer a priori definitions than many previously used methods. Thus future studies on job postings may adopt our method. Furthermore, the method can also be applied to other online curriculum databases where curriculum texts are available to the international research audience, such as e.g., the Open Syllabus Project at Columbia University, which collects millions of digital curricula in its database and makes them available to the public.

Moreover, our results lead to recommendations for policy makers who want to prepare workers for the increasing diffusion of IT. The policy recommendations apply for dual apprenticeship systems as well as systems implementing work-related training such as the UK (see e.g., Fettes, Evans, and Kashefpakdel 2020, who stress the importance of 'putting skills to work' and how context matters). Specifically, we show results regarding the success of different IT trainings as indicated by labour market outcomes. We argue that when deciding on policy changes (for example curricula designs), policy makers should recognize the importance of skills packages with IT and what they may look like. A successful IT training consists of skills packages with IT and its complementary set of skills that are typically used in the workplace. For example, a skills package with IT-applications should also include, e.g., project management, which helps improve the efficiency of any type of IT-application and thus elevates the productivity of the single IT skill.

References

Amado, Alexandra, Paulo Cortez, Paulo Rita, and Sérgio Moro. 2018. "Research Trends on Big Data in Marketing: A Text Mining and Topic Modeling Based Literature Analysis." *European Research on Management and Business Economics* 24 (1): 1–7. doi:10.1016/j.iedeen.2017.06.002.

Ambrosino, Angela, Mario Cedrini, John B. Davis, Stefano Fiori, Marco Guerzoni, and Massimiliano Nuccio. 2018. "What Topic Modeling Could Reveal About the Evolution of Economics." *Journal of Economic Methodology* 25 (4): 329–48. doi:10.1080/1350178X.2018.1529215.

Angrist, Joshua, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton Univ. Press.

Atalay, Enghin, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. 2018. "New Technologies and the Labor Market." *Journal of Monetary Economics* 97:48–67. doi:10.1016/j.jmoneco.2018.05.008.

———. 2020. "The Evolution of Work in the United States." *American Economic Journal: Applied Economics*, no. 2: 1–34. doi:10.1257/app.20190070.

Backes-Gellner, Uschi, and Regula Geel. 2014. "A Comparison of Career Success Between Graduates of Vocational and Academic Tertiary Education." *Oxford Review of Education* 40 (2): 266–91. doi:10.1080/03054985.2014.889602.

Balestra, Simone, and Uschi Backes-Gellner. 2017. "When a Door Closes, a Window Opens? Long-Term Labor Market Effects of Involuntary Separations." *German Economic Review* 18 (1): 1–21. doi:10.1111/geer.12086.

Benites-Lazaro, L. L., L. Giatti, and A. Giarolla. 2018. "Topic Modeling Method for Analyzing Social Actor Discourses on Climate Change, Energy and Food Security." *Energy Research & Social Science* 45:318–30. doi:10.1016/j.erss.2018.07.031.

Borghans, Lex, and Bas ter Weel. 2006. "Do We Need Computer Skills to Use a Computer? Evidence from Britain." *Labour* 20 (3): 505–32. doi:10.1111/j.1467-9914.2006.00351.x.

Brown, Phillip, and Manuel Souto-Otero. 2020. "The End of the Credential Society? An Analysis of the Relationship Between Education and the Labour Market Using Big Data." *Journal of Education Policy* 35 (1): 95–118. doi:10.1080/02680939.2018.1549752.

Buchmann, Marlis, Helen Buchs, and Ann-Sophie Gnehm. 2020. "Occupational Inequality in Wage Returns to Employer Demand for Types of Information and Communications Technology (ICT) Skills: 1991–2017." *Köln Z Soziol* 72 (S1): 455–82. doi:10.1007/s11577-020-00672-5.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information Processing Systems 22*, edited by Yoshua Bengio, A. Culotta, J. D. Lafferty, D. Schuurmans, and C. K. I. Williams, 288–96: Curran Associates, Inc.

Deming, David, and Lisa B. Kahn. 2017. “Skill Requirements Across Firms and Labor Markets: Evidence from Job Postings for Professionals.” *Journal of Labor Economics* 36 (S1): S337-S369. doi:10.1086/694106.

DiNardo, John E., and Jörn-Steffen Pischke. 1997. “The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?” *QJE* 112 (1): 291–303. http://www.jstor.org/stable/2951283.

Eggenberger, Christian, and Uschi Backes-Gellner. 2020. “IT Skills, Occupation Specificity and Job Separations.” Swiss Leading House "Economics of Education" Working Paper 172.

Eggenberger, Christian, Miriam Rinawi, and Uschi Backes-Gellner. 2018. “Occupational Specificity: A New Measurement Based on Training Curricula and Its Effect on Labor Market Outcomes.” *Labour Economics* 51:97–107. doi:10.1016/j.labeco.2017.11.010.

Falck, Oliver, Alexandra Heimisch-Roecker, and Simon Wiederhold. 2020. “Returns to ICT Skills.” *Research Policy*, 104064. doi:10.1016/j.respol.2020.104064.

Federal Statistical Office. 2011. “Syntheseerhebung Soziale Sicherheit Und Arbeitsmarkt (SESAM): Grundlagen, Methoden, Konstruierte Variablen.” https://www.bfs.admin.ch/bfs/de/home/statistiken/arbeit-erwerb/erhebungen/sesam.assetdetail.322180.html.

Fettes, Trisha, Karen Evans, and Elnaz Kashefpakdel. 2020. “Putting Skills to Work: It’s Not so Much the What, or Even the Why, but How…”. *Journal of Education and Work* 33 (2): 184–96. doi:10.1080/13639080.2020.1737320.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. 2019. “Text as Data.” *Journal of Economic Literature* 57 (3): 535–74. doi:10.1257/jel.20181020.

Gillis, Nicolas. 2014. “The Why and How of Nonnegative Matrix Factorization.” Chapman & Hall / CRC machine learning & pattern recognition series. http://arxiv.org/pdf/1401.5226v2.

Goetze, Walter, and Birgül Aksu. 2018. “Anforderungsprofile: Die Methode.” Accessed April 21, 2020. http://www.anforderungsprofile.ch/index.cfm?action=act_getfile&doc_id=100052&.

Grimmer, Justin, and Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21 (3). doi:10.1093/pan/mps028.

Grinis, Inna. 2019. “The STEM Requirements of “Non-STEM” Jobs: Evidence from UK Online Vacancy Postings.” *Economics of Education Review* 70:144–58. doi:10.1016/j.econedurev.2019.02.005.

Jaik, Katharina, and Stefan C. Wolter. 2019. “From Dreams to Reality: Market Forces and Changes from Occupational Intention to Occupational Choice.” *Journal of Education and Work* 32 (4): 320–34. doi:10.1080/13639080.2019.1637830.

Jansen, Anika, Andries de Grip, and Ben Kriechel. 2017. “The Effect of Choice Options in Training Curricula on the Demand for and Supply of Apprentices.” *Economics of Education Review* 57: 52–65. https://doi.org/10.1016/j.econedurev.2017.02.003.

Janssen, Simon, and Jens Mohrenweiser. 2018. “The Shelf Life of Incumbent Workers During Accelerating Technological Change: Evidence from a Training Regulation Reform.” IZA Discussion Papers 11312.

Kiener, Fabienne, Ann-Sophie Gnehm, and Uschi Backes-Gellner. 2020. “Non-Cognitive Skills in Training Curricula and Heterogenous Wage Returns.” Swiss Leading House "Economics of Education" Working Paper 175.

Krueger, A. B. 1993. “How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984-1989.” *QJE* 108 (1): 33–60. doi:10.2307/2118494.

Lazear, Edward P. 2009. “Firm-specific Human Capital: A Skill-weights Approach.” *Journal of Political Economy* 117 (5): 914–40. doi:10.1086/648671.

Lee, Daniel D., and H. Sebastian Seung. 1999. “Learning the Parts of Objects by Non-Negative Matrix Factorization.” *Nature* 401 (6755): 788–91. doi:10.1038/44565.

Maier, Daniel, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch et al. 2018. “Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology.” *Communication Methods and Measures* 12 (2-3): 93–118. doi:10.1080/19312458.2018.1430754.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” https://arxiv.org/pdf/1301.3781.

Quinn, Kevin, Burt L. Monroe, Michael Colaresi, Michael Crespin, and Dragomir Radev. 2010. “How to Analyze Political Text with Minimal Assumptions and Costs.” *American Journal of Political Science* 54 (1). doi:10.1111/j.1540-5907.2009.00427.x.

Reisenbichler, Martin, and Thomas Reutterer. 2019. “Topic Modeling in Marketing: Recent Advances and Research Opportunities.” *Journal of Business Economics* 89 (3): 327–56. doi:10.1007/s11573-018-0915-7.

Schultheiss, Tobias, Curdin Pfister, Uschi Backes-Gellner, and Ann-Sophie Gnehm. 2018. “Tertiary Education Expansion and Task Demand: Does a Rising Tide Lift All Boats?” Swiss Leading House "Economics of Education" Working Paper 154.

Shahnaz, Farial, Michael W. Berry, V.Paul Pauca, and Robert J. Plemmons. 2006. “Document Clustering Using Nonnegative Matrix Factorization.” *Information Processing & Management* 42 (2): 373–86. doi:10.1016/j.ipm.2004.11.005.

Shields, Robin, and Andres Sandoval Hernandez. 2020. “Mixed Signals: Cognitive Skills, Qualifications and Earnings in an International Comparative Perspective.” *Oxford Review of Education* 46 (1): 111–28. doi:10.1080/03054985.2019.1687436.

Spöttl, Georg, and Lars Windelband. 2021. “The 4 Th Industrial Revolution – Its Impact on Vocational Skills.” *Journal of Education and Work* 34 (1): 29–52. doi:10.1080/13639080.2020.1858230.

Tambe, Prasanna, Xuan Ye, and Peter Cappelli. 2020. “Paying to Program? Engineering Brand and High-Tech Wages.” *Management Science* 66 (7): 3010–28. doi:10.1287/mnsc.2019.3343.

Tjioe, Elina, Michael Berry, Ramin Homayouni, and Kevin Heinrich. 2008. “Using a Literature-Based NMF Model for Discovering Gene Functional Relationships.” *BMC Bioinformatics* 9 (Suppl 7): P1. doi:10.1186/1471-2105-9-S7-P1.

Wolter, Stefan C., and Maria Zumbuehl. 2017. “The Native-Migrant Gap in the Progression into and Through Upper-Secondary Education.” CESifo Working Paper Series 6810.

Yan, Xiaohui, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang. 2013. “Learning Topics in Short Texts by Non-Negative Matrix Factorization on Term Correlation Matrix.” In *Proceedings of the 2013 SIAM International Conference on Data Mining*, edited by Joydeep Gosh, 749–57. [Philadelphia, PA]: SIAM, Society for Industrial and Applied Mathematics.

Appendix A: Natural language processing methods

A1. Preparing the texts

We prepare the curriculum texts according to NLP standards. Words without meaningful content like articles ('the', 'their') are excluded. Furthermore, the words are stemmed, such that different forms like plurals or verb conjunctions possess the same stem.

The text is represented not as single expressions of words but as bag of words that contain several words that occur next to each other. The advantage of a bag-of-words representation is that it considers the previous or following word (i.e., context), which is needed because in some cases an expression has only the correct meaning when more than one word is displayed (e.g., 'communication system'). The researcher can choose how many words the bag should contain (uni-, bi-, trigram). We include single words (unigrams), two words together (bigrams) and three words together (trigrams). For example, the sentence 'apprentice develops user-friendly applications' is in a trigram representation {apprentice, develops, user-friendly} or in a bigram representation {the, apprentice}, {apprentice, develops}, {develops, user-friendly}, {user-friendly, applications} or in a unigram representation {the}, {apprentice}, {develops}, {user-friendly}, {applications}. The sequence of the words within a bag-of-word representation does not matter, meaning that a bag of words {apprentice, develops} is the same as {develops, apprentice}.

A2. Illustrative examples of the matrix factorization

From one input matrix, for example many text documents such as our curricula, NMF generates two output matrices, 'Words by Topics', which shows the words important for each topic, and 'Topics by Documents', which shows the distribution of the topics over the documents. Thus the algorithm factorizes the input matrix into the output matrices.

The input matrix shows the specific words in each curriculum, considering not only single words but also words occurring next to another (see the previous subsection on the preparation of the texts). Each word in each curriculum is attached with a value for the relative importance of the word for that curriculum (a process called 'term frequency-inverse document frequency', tf-idf). In other words, the value represents how important the word in the curriculum is relative to all words occurring in the curriculum times a weight of how

important the word is over all curricula. We illustrate the input matrix with a fictional example.

Table A1. Illustrative, partly fictional input matrix.

Words / *Curricula (Document)*	*{apprentice, develops}*	*{develops, user-friendly}*	*{user-friendly, applications}*
Information Technologist	0.1	0.3	0.2
Photographer	0.4	0	0

Notes: Authors' illustrative, partly fictional example of the input matrix.

Suppose a curriculum contains the sentence 'apprentice develops user-friendly applications.' For the input matrix, the sentence is split in pairs of words: apprentice, develops; develops, user-friendly; user-friendly, applications. If the words do not appear in a curriculum, the tf-idf value is 0. In the example, each word (pair) is important in the curriculum of the information technologist, whereas in the curriculum of the photographer only 'apprentice develops' occurs.

From the input matrix, the algorithm builds two output matrices. The algorithm learns by itself how to construct the output matrices such that they optimally represent the input matrix. One output matrix, Words by Topic, shows the most important words for each topic. The algorithm determines the most important words for each topic and attaches a weight to each word representing the importance. The weight is a positive value below one. An illustrative output matrix Words by Topic may look as follows:

Table A2. Illustrative, partly fictional output matrix words by topic.

Topic	*Most important words*
Topic 1	apprentice develops (0.28), user-friendly application (0.14)
Topic 2	image editing (0.292), …

Notes: Authors' illustrative, partly fictional example of the output matrix Words by Topics. In parentheses weights showing the relative importance of the words by topic.

The second output matrix, Topics by Document, shows the distribution of the topics across each curriculum. It also attaches a weight how important each topic in each curriculum (also a value between 0 and 1). The illustrative example may look as follows:

Table A3. Illustrative, partly fictional output matrix topics by document.

Curriculum (Document)	*Topic 1*	*Topic 2*
Information Technologist	0.2	0.01
Photographer	0.1	0.3

Notes: Authors' illustrative, partly fictional example of the output matrix Topics by Document. Each cell shows a weight representing the importance of the topic in the curriculum.

A3. Interpretation tool

To find IT words in the skills packages, we apply an 'interpretation tool.' To start, we—the researchers—need to determine some initial IT words from the skills packages, which then are extended through an automatic interpretation tool. The procedure follows four steps: First, we pick some initial IT words that are clearly IT-related, for example 'automation,' 'CNC,' 'CAD,' 'to configure,' 'ICT,' and 'application.' Second, the interpretation tool detects all similar words in the corpus.[29] Third, the interpretation tool compares the similar words to our topic-modelling output words by topic. For example, when we input 'to configure', the tool finds 'to install' or 'to test' as word similarities. Fourth, if a word in the topic modelling output is similar to an IT word, then the tool categorizes it as an IT word. We then have a comprehensive set of IT vocabulary.

Resulting IT vocabulary in words by topic *(input words by researchers in italics):*
application, application development, explain application development, explain application development independently, *automate, user, user terminal device, user, user user, user interfaces, business informatics,* explain business informatics, business informatics explain independently, *cad, cam, cnc, database,* data model, privacy, electrical system technologies, *develop,* production technologies, production technologies machine technologies, production technologies machine technologies control engineering, building automation, *ict,* implement, informatics resources, installation, install, installing, realize

[29] We build 100-dimensional word vectors, with the word2vec CBOW algorithm and a window size of five, excluding words with a frequency lower than five. The model was trained over 20 iterations. Similar words refer to words with a cosine similarity of 0.6 or higher, at the most 100 suggestions are accepted.

small projects, *communication system, configure,* course application development, course explain application development, course business informatics, explaining the course in business informatics, apprentices develop, apprentices install, machine technologies, machine technologies control technologies, *net, network infrastructure, network, pc, use pc,* practice course application development, practice course business informatics, realize, *restore, server, server service, software*, system, *system technologies,* explain system technologies, telematics systems, test, testing, realize projects, *web presence*

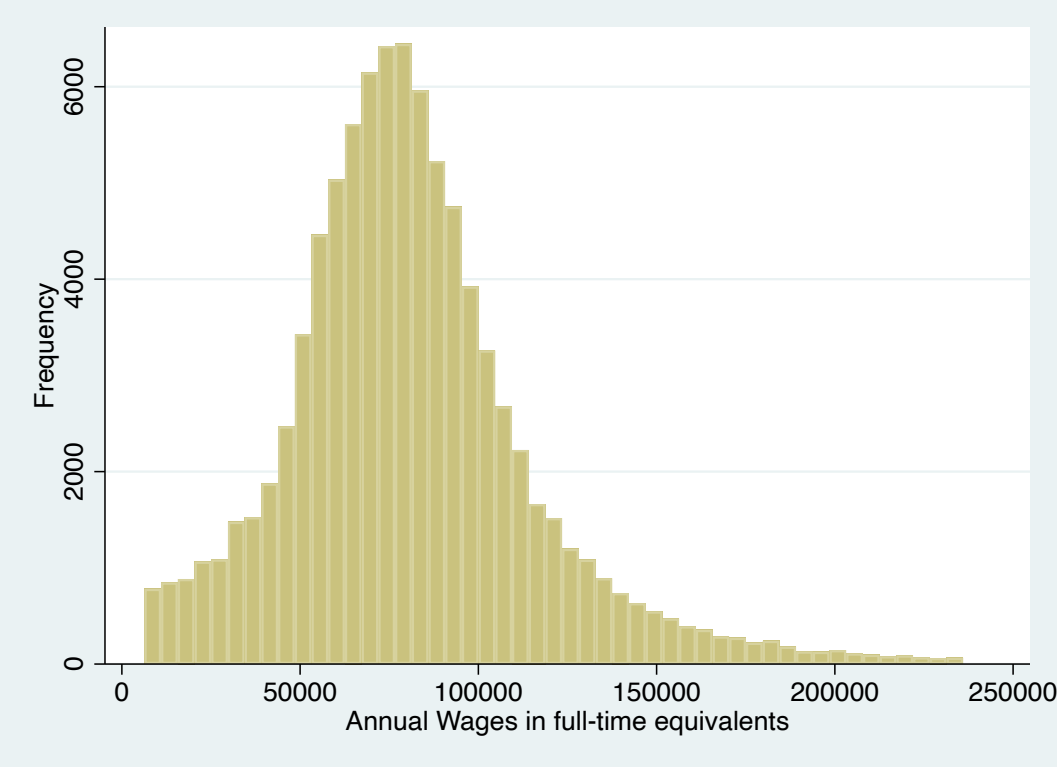
Appendix B: Additional tables

Table B1. Skills packages without IT: Matrix words by topics.

Topic	*Relevant words*
Topic 1	act, competent, according to, following, standard, environmental protection
Topic 2	agriculture, plant cultivation, be in force, organic farming
Topic 4	course, to name, technical building system, building technical, system
Topic 5	explain, education and training, vocational education and training
Topic 6	process-oriented economically, process-oriented economic thinking, act think economically
Topic 7	explain understand, apply, name know, describe
Topic 8	explain, act, competent, explain, guest, course, meaning
Topic 11	situation, public, explain, benchmark requirement, customer, correct, comprehensible

Notes: Excerpt of output matrix Words by Topics resulting from NMF of the curriculum database. Not included are topic 3, 9, 10, and 12 since they are skills packages with IT.

Figure B1. Histogram of wages: Wage sample.



Notes: Authors' calculations of annual wages in sample, data based on the SESAM, 2010–2016.

Table B2. Joint F-test for each pair of coefficients, main wage analysis, table 3, regression (5).

	Skills package w/ CNC/CAD	*Skills package w/ system technologies*	*Skills package w/ IT-applications*	*Skills package w/ control technologies*
Skills package w/ CNC/CAD				
Skills package w/ system technologies	F-test = 4.96 p-value = 0.0260			
Skills package w/ IT-applications	F-test = 36.67 p-value = 0.0000	F-test = 127.72 p-value = 0.0000		
Skills package w/ control technologies	F-test = 3.10 p-value = 0.0778	F-test = 5.96 p-value = 0.0146	F-test = 0.09 p-value = 0.7601	

Notes: Authors' calculations of joint F-test for each pair of coefficients of regression (5) in table 3. Data based on their skills measures and the SESAM, 2010–2016.

Table B3. Wages and skills packages with IT: Analysis adding controls one by one.

VARIABLES	(1) *log wage*	(2) *log wage*	(3) *log wage*	(4) *log wage*	(5) *log wage*
Skills package w/ CNC/CAD	0.050***	0.045***	0.027***	0.027***	0.027***
	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)
Skills package w/ system technologies	0.029***	0.034***	0.019***	0.019***	0.019***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Skills package w/ IT-applications	0.065***	0.066***	0.059***	0.058***	0.058***
	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
Skills package w/ control technologies	0.044***	0.041***	0.054***	0.054***	0.054***
	(0.014)	(0.014)	(0.015)	(0.015)	(0.015)
age, age²		Yes	Yes	Yes	Yes
gender			Yes	Yes	Yes
Swiss				Yes	Yes
years					Yes
Constant	11.177***	9.660***	9.737***	9.696***	9.671***
	(0.018)	(0.043)	(0.053)	(0.058)	(0.060)
Observations	89,259	89,259	89,259	89,259	89,259
Number of individuals	56,014	56,014	56,014	56,014	56,014
R^2 overall	0.0339	0.116	0.142	0.144	0.145
R^2 between individuals	0.0335	0.121	0.147	0.149	0.150

Notes: Authors' calculations, based on their skills measures and the SESAM, 2010–2016. Robust standard errors in parentheses clustered on training occupation. *** p<0.01, ** p<0.05, * p<0.1. Regression (5) is the same as regression (5) in table 3.

Table B4. Wages and skills packages with IT: Analysis controlling for dummies for skills packages without IT.

VARIABLES	(1) *log wage*	(2) *log wage*
Skills package w/ CNC/CAD	0.027***	0.050***
	(0.007)	(0.016)
Skills package w/ system technologies	0.019***	0.039***
	(0.005)	(0.013)
Skills package w/ IT-applications	0.058***	0.062***
	(0.005)	(0.009)
Skills package w/ control technologies	0.054***	0.035**
	(0.015)	(0.017)
age, age²	Yes	Yes
gender	Yes	Yes
Swiss	Yes	Yes
years	Yes	Yes
dummies for skills packages without IT		Yes
Constant	9.671***	9.575***
	(0.060)	(0.084)
Observations	89,259	89,259
Number of individuals	56,014	56,014
R^2 overall	0.145	0.156
R^2 between individuals	0.150	0.161

Notes: Authors' calculations, based on their skills measures and the SESAM, 2010–2016. Robust standard errors in parentheses clustered on training occupation. *** p<0.01, ** p<0.05, * p<0.1. Regression (1) is the same as regression (5) in table 3 or regression (5) in table B3.

Table B5. Joint F-test for each pair of coefficients, main employment analysis, table 5, regression (5).

	Skills package w/ CNC/CAD	*Skills package w/ system technologies*	*Skills package w/ IT-applications*	*Skills package w/ control technologies*
Skills package w/ CNC/CAD				
Skills package w/ system technologies	F-test = 13.13 p-value = 0.0003			
Skills package w/ IT-applications	F-test = 17.96 p-value = 0.0000	F-test = 7.83 p-value = 0.0051		
Skills package w/ control technologies	F-test= 1.91 p-value = 0.1671	F-test = 0.06 p-value = 0.8098	F-test = 0.23 p-value = 0.6333	

Notes: Authors' calculations of joint F-test for each pair of coefficients of regression (5) in table 5. Data based on their skills measures and the Labour Market Survey (SESAM), 2010–2016.

Table B6. Employment and skills packages with IT: Adding controls one by one.

VARIABLES	(1) *employed*	(2) *employed*	(3) *employed*	(4) *employed*	(5) *employed*
Skills package w/ CNC/CAD	0.010***	0.012***	0.012***	0.012***	0.012***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Skills package w/ system technologies	0.008***	0.008***	0.008***	0.008***	0.008***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Skills package w/ IT-applications	0.007***	0.006***	0.006***	0.006***	0.005**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Skills package w/ control technologies	0.006*	0.007**	0.007*	0.007**	0.007*
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
age, age²		Yes	Yes	Yes	Yes
gender			Yes	Yes	Yes
Swiss				Yes	Yes
years					Yes
Constant	0.921***	0.925***	0.926***	0.931***	0.930***
	(0.007)	(0.018)	(0.019)	(0.021)	(0.019)
Observations	99,822	99,822	99,822	99,822	99,822
Number of individuals	61,752	61,752	61,752	61,752	61,752
R^2 overall	0.00322	0.0112	0.0112	0.0113	0.0125
R^2 between individuals	0.00323	0.0131	0.0131	0.0133	0.0146

Notes: Authors' calculations, based on their skills measures and the SESAM, 2010–2016. Robust standard errors in parentheses clustered on training occupation. Regression (5) is the same as regression (5) in table 5. *** p<0.01, ** p<0.05, * p<0.1.

Table B7. Employment and skills packages with IT: Controlling for dummies for skills packages without IT.

VARIABLES	(1) *employed*	(2) *employed*
Skills package w/ CNC/CAD	0.012***	0.020***
	(0.003)	(0.008)
Skills package w/ system technologies	0.008***	0.015**
	(0.002)	(0.006)
Skills package w/ IT-applications	0.005**	0.008**
	(0.002)	(0.004)
Skills package w/ control technologies	0.007*	0.004
	(0.004)	(0.006)
age, age²	Yes	Yes
gender	Yes	Yes
Swiss	Yes	Yes
years	Yes	Yes
dummies for skills packages without IT		Yes
Constant	0.930***	0.891***
	(0.019)	(0.031)
Observations	99,822	99,822
Number of individuals	61,752	61,752
R^2 overall	0.0125	0.0178
R^2 between individuals	0.0146	0.0199

Notes: Authors' calculations, based on their skills measures and the SESAM, 2010–2016. Robust standard errors in parentheses clustered on training occupation. Regression (1) is the same as regression (5) in table 5 or regression (5) in table B6. *** p<0.01, ** p<0.05, * p<0.1.